STAT525 HOMEWORK#1

1. KNNL Problem 1.26

2. A regression analysis of a set of data produced the following fitted equation: $\hat{y} = 3+8x$.

   (a) If x increases 5 units, how does $\hat{y}$ change?

   (b) Here x was measured in degrees Celsius. Rewrite the fitted equation with x replaced by x^* where x^* is x expressed in degrees Fahrenheit. Use the fact that $x = (5/9) \times (x^* - 32)$.

3. KNNL Problem 1.39 part a.

   Hint $(Y_1 - a)^2 + (Y_2 - a)^2 = 2(\bar{Y} - a)^2 + (Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2$, where $\bar{Y} = (Y_1 + Y_2)/2$ for any Y_1, Y_2 and a.

4. Derive the MLE estimator for $(\beta_0, \beta_1, \sigma^2)$ for simple linear regression with normal error.

5. Show that $s^2 = \sum(Y_i - \hat{Y}_i)^2/(n-2)$ is an unbiased estimator for σ^2.

1. KNNL Problem 1.26

Plastic hardness. Refer to Problems 1.3 and 1.14. Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours, and Y is hardness in Brinell units. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	14	15	16
X_i:	16	16	16	...	40	40	40
Y_i:	199	205	196	...	248	253	246

1.26. Refer to **Plastic hardness** Problem 1.22.

a. Obtain the residuals e_i. Do they sum to zero in accord with (1.17)?

```
DATA a1;
   INPUT y x;
   CARDS;
199 16
205 16
196 16
200 16
218 24
220 24
215 24
223 24
237 32
234 32
235 32
230 32
250 40
248 40
253 40
246 40
;
RUN;

PROC PRINT DATA=a1; RUN;
```

```
/* Regression, residuals, predictions */
PROC REG DATA=a1;
   MODEL y = x / CLB P R;
   OUTPUT OUT=a2 P=pred R=resid;
   ID x;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y

Output Statistics

Obs	x	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D
1	16	199	201.1500	1.3529	-2.1500	2.937	-0.732	0.057
2	16	205	201.1500	1.3529	3.8500	2.937	1.311	0.182
3	16	196	201.1500	1.3529	-5.1500	2.937	-1.753	0.326
4	16	200	201.1500	1.3529	-1.1500	2.937	-0.391	0.016
5	24	218	217.4250	0.8857	0.5750	3.110	0.185	0.001
6	24	220	217.4250	0.8857	2.5750	3.110	0.828	0.028
7	24	215	217.4250	0.8857	-2.4250	3.110	-0.780	0.025
8	24	223	217.4250	0.8857	5.5750	3.110	1.792	0.130
9	32	237	233.7000	0.8857	3.3000	3.110	1.061	0.046
10	32	234	233.7000	0.8857	0.3000	3.110	0.096	0.000
11	32	235	233.7000	0.8857	1.3000	3.110	0.418	0.007
12	32	230	233.7000	0.8857	-3.7000	3.110	-1.190	0.057
13	40	250	249.9750	1.3529	0.0250	2.937	0.009	0.000
14	40	248	249.9750	1.3529	-1.9750	2.937	-0.672	0.048
15	40	253	249.9750	1.3529	3.0250	2.937	1.030	0.112
16	40	246	249.9750	1.3529	-3.9750	2.937	-1.353	0.194

Sums

$X=16$: $-2.150, 3.850, -5.150, -1.150 \Rightarrow -4.6$
$X=24$: $0.575, 2.575, -2.425, 5.575 \Rightarrow 6.3$
$X=32$: $3.300, 0.300, 1.300, -3.700 \Rightarrow 1.2$
$X=40$: $0.025, -1.975, 3.025, -3.975 \Rightarrow -2.9$

Grand Sums $\Rightarrow 0$

Yes, Residuals do sum up to 0, confirming property (1.17).

b. Estimate σ^2 and σ. In what units is σ expressed?

$\boxed{\sigma^2 = MSE}$ $\boxed{MSE = SSE/df}$

- $SSE = 146.425$
- $MSE = 10.459$
- $df_{Error} = 16-2 = 14$

$\hat{\sigma}^2 = 10.459$

$\Downarrow$

Units: $\hat{\sigma}^2$ is hardness^2, Y^2 (plastic hardness units)

$\hat{\sigma} = \sqrt{10.459} = 3.234$

$\Downarrow$

Units: $\hat{\sigma}$ has the same unit as Y (plastic hardness units)

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5297.51250	5297.51250	506.51	<.0001
Error	14	146.42500	10.45893		
Corrected Total	15	5443.93750			

Root MSE	3.23403	R-Square	0.9731
Dependent Mean	225.56250	Adj R-Sq	0.9712
Coeff Var	1.43376		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	168.60000	2.65702	63.45	<.0001	162.90125	174.29875
x	1	2.03438	0.09039	22.51	<.0001	1.84050	2.22825

2. A regression analysis of a set of data produced the following fitted equation: $\hat{y} = 3 + 8x$.

(a) If x increases 5 units, how does $\hat{y}$ change?

$\hat{y} = 3 + \underbrace{8(5)} = 43$

$\Delta\hat{y} = 40$ $\hat{y}$ increases by 40 units.

(b) Here x was measured in degrees Celsius. Rewrite the fitted equation with x replaced by x^* where x^* is x expressed in degrees Fahrenheit. Use the fact that $x = (5/9) \times (x^* - 32)$.

$\hat{y} = 3 + 8(5/9(x^* - 32))$

$= 3 + 40/9\,(x^* - 32))$

$= 3 + 40/9\,x^* - 1280/9$

$= 40/9\,x^* + 27/9 - 1280/9$

$\hat{y} = 40/9\,x^* - 1253/9$ or $\hat{y} = 4.44\,x^* - 139.22$

- 6 observations: x2 Y's at each $X = 5, 10, 15$

- Y_1 and Y_2 are two observations at the same X level
- $\bar{Y} = (Y_1 + Y_2)/2$ is the mean.
- 'a' is the target value you want to compare them to.

↓

fitted value from the regression line at line X:

$a = \hat{Y}(X)$

- $2(\bar{Y} - a)^2$: depend on mean of replicates + regression line to derive a_i.
- $(Y_i - \bar{Y})^2$: depend on deviations within replicates ⇒ constant

↳ ●: within cell, do not involve a, only $\bar{Y}$ matter.
($\hat{Y}$ line)

splitting sum of squares into 2 parts

3. KNNL Problem 1.39 part a.

Hint $(Y_1 - a)^2 + (Y_2 - a)^2 = \boxed{2(\bar{Y} - a)^2} + \boxed{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2}$, where $\bar{Y} = (Y_1 + Y_2)/2$ for any Y_1, Y_2 and a.

1.39. Two observations on Y were obtained at each of three X levels, namely, at $X = 5$, $X = 10$, and $X = 15$.

a. Show that the least squares (OLS) regression line fitted to the *three* points $(5, \bar{Y}_1)$, $(10, \bar{Y}_2)$, and $(15, \bar{Y}_3)$, where $\bar{Y}_1$, $\bar{Y}_2$, and $\bar{Y}_3$ denote the means of the Y observations at the three X levels, is identical to the least squares regression line fitted to the original six cases.

- We can 2 observations on Y at each level X ($X = 5, 10, 15$).
- Let these be Y_{j1}, Y_{j2} at level $X = x_j$, and let $\bar{Y}_j = \frac{Y_{j1} + Y_{j2}}{2}$

- The LS Regression minimizes:

3xs 2obs each X

$$SSE = \sum_{j=1}^{3} \sum_{i=1}^{2} (Y_{ij} - \hat{Y}(x_j))^2$$

- Step 1. Apply the hint

- For any 2 obs. Y_1, Y_2, with mean $\bar{Y}$:

$$(Y_1 - a)^2 + (Y_2 - a)^2 = 2(\bar{Y} - a)^2 + (Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2$$

$$\Downarrow$$

$$(Y_{j1} - \hat{Y}(x_j)) + (Y_{j2} - \hat{Y}(x_j))^2 = 2(\bar{Y}_j - \hat{Y}(x_j))^2 + (Y_{j1} - \bar{Y}_j)^2 + (Y_{j2} - \bar{Y}_j)^2$$

- Step 2. Substitude into SSE:

Depend on $\hat{Y}$ so it matters when minimizing SSE.

$$SSE = \sum_{j=1}^{3} \Big[2(\bar{Y}_j - \hat{Y}(x_j))^2 + \underbrace{(Y_{j1} - \bar{Y}_j)^2 + (Y_{j2} - \bar{Y}_j)^2}_{\text{only depend on data, not the regression line, } \hat{Y} \text{ does not matter}} \Big]$$

- Step 3. Simplify the minimization problem.

weight, n_j

$$SSE = 2 \sum_{j=1}^{3} [\bar{Y}_j - \hat{Y}(x_j)]^2 + \text{Constant} \longrightarrow b_1 = \frac{\sum_j \cancel{n_j} (x_i - \bar{x}_w)(\bar{Y}_j - \bar{y}_w)}{\sum_j \cancel{n_j} (x_j - \bar{x}_w)^2}$$

number of observations in each x_j level

- This is the same as fitting the regression line to the 3 mean points $(5, \bar{Y}_1)$ $(10, \bar{Y}_2)$ $(15, \bar{Y}_3)$, each with weight $2 = n_j$
- Since the weights are equal, the line is the same as the 'unweighted' regression on just the 3 means.
  ↳ exactly 2 obs. at each level

Therefore, the LS Regression line fitted to the 6 original cases is identical to the line fitted to the 3 mean points.

4. Derive the MLE estimator for $(\beta_0, \beta_1, \sigma^2)$ for simple linear regression with normal error.

error (residuals) in the LR Model follow normal distribution

- We assume the model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \theta^2)$
$Y_i \sim N(\beta_0 + \beta_1 x_i, \theta^2)$

- Step 1. Likelihood Function:

Because the errors are independent normal, the joint density is:

$$L(\beta_0, \beta_1, \theta^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta^2}} \exp\left[-\frac{1}{2\theta^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

- Step 2. Take the Log-Likelihood

$$\ell(\beta_0, \beta_1, \theta^2) = \log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\theta^2) \underbrace{- \frac{1}{2\theta^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}_{\text{depend on } \beta_0, \beta_1, \theta^2}$$

- Step 3. Maximize the Log-Likelihood (β_0, β_1)

To maximize ℓ, we minimize:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \quad \text{(same as OLS)}$$

- Partial derivative for β_0:

$$\partial/\partial\beta_0 \; SSE = -2\sum(y_i - \beta_0 - \beta_1 x_i)$$
$\Downarrow$ set to 0
$$\sum y_i - n\beta_0 + \beta_1 \sum x_i = 0 \quad (1)$$
$\Downarrow$ rewrite
$$n\bar{y} = n\beta_0 + \beta_1 n\bar{x}$$
$\Downarrow$ solve
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- Solve normal equations:

- Let $\bar{x} = \frac{1}{n}\sum x_i$
$\bar{y} = \frac{1}{n}\sum y_i$

- Partial derivative for β_0:

$$\partial/\partial\beta_1 \; SSE = -2\sum(y_i - \beta_0 - \beta_1 x_i)$$
$\Downarrow$ set to 0
$$\sum x_i y_i - \beta_0 \sum x_i - \beta_1 \sum x_i^2 = 0 \quad (2)$$
$\Downarrow$ Plug in $\beta_0 = \bar{y} - \beta_1\bar{x}$ and solve:
$$\sum x_i y_i - (\bar{y} - \beta_1\bar{x})\sum x_i - \beta_1 \sum x_i^2 = 0$$
$$\sum x_i y_i - \bar{y}\sum x_i + \beta_1\bar{x}\sum x_i - \beta_1\sum x_i^2 = 0$$
$$\beta_1\left(\sum x_i\bar{x} - \sum x_i^2\right) = \bar{y}\sum x_i - \sum x_i y_i$$
$\Downarrow$ simplify
$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad \text{(same as OLS)}$$

- Step 4. Maximize θ^2

Plug $\hat{\beta}_0$ and $\hat{\beta}_1$ back into Log-Likelihood:

$$\ell(\beta_0, \beta_1, \theta^2) = (\text{Constant}) - \frac{n}{2}\log(\theta^2) - \frac{1}{2\theta^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Differentiate θ^2, set to 0:

$$\frac{d\ell}{d\theta^2} = -\frac{n}{2\theta^2} + \frac{1}{2(\theta^2)^2}\sum e_i^2 = 0$$
$\Downarrow$ solve
$$\hat{\theta}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{SSE}{n} \quad \text{biased}$$

(OLS uses $\frac{1}{n-2}$) unbiased

$\downarrow$

$$\boxed{\begin{aligned} \hat{\beta}_1 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\theta}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned}}$$

5. Show that $s^2 = \sum(Y_i - \hat{Y}_i)^2/(n-2)$ is an unbiased estimator for σ^2.

Step 1. Define Error Model

In simple linear regression, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$

The fitted value, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

The residual, $e_i = Y_i - \hat{Y}_i$

Step 2. Define SSE and s^2

$$SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$$

$$s^2 = SSE / n-2$$

we want to show: $E(s^2) = \sigma^2$

2 df lost by using $\hat{\beta}_0, \hat{\beta}_1$ in place of β_0, β_1

Step 3. Use Theorem:

If the model is correct and errors are independent $\sim N(0, \sigma^2)$, then,

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-2}$$

so,

$$E(SSE) = (n-2)\sigma^2$$

$$\Downarrow$$

$$E(s^2) = \frac{1}{n-2} \cdot E(SSE) = \sigma^2$$

Hence,

$E(s^2) = \sigma^2 \rightarrow s^2$ is an unbiased estimator of σ^2.

The MLE divides by n because it maximizes the likelihood without adjusting for parameter estimation, while the LSE divides by $n-2$ to correct for the loss of two degrees of freedom from estimating β_0 and β_1, making it an unbiased estimator.