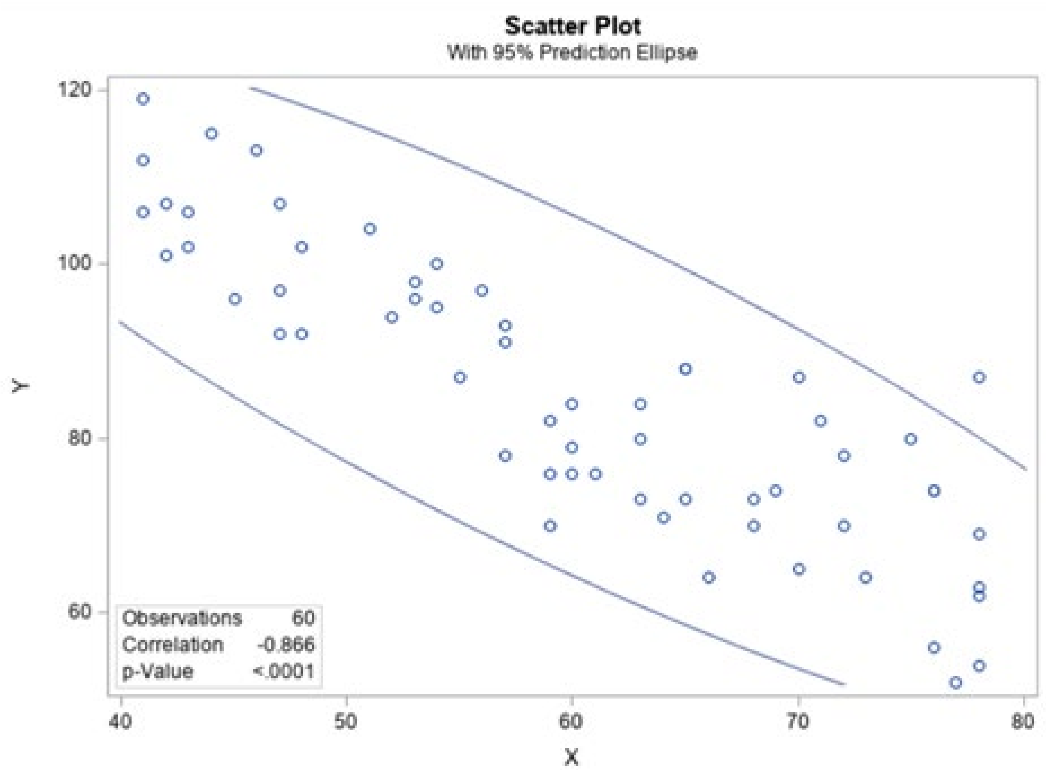
Adv Regression, Homework 1

Yiwen Ren

1.



Comments: From the graph above, we can see there is a negative linear association between age and muscle mass, i.e., with the age increasing, the muscle mass decreased.

(b) Yes, there is a linear association between Y and X.

Pearson Correlation Statistics (Fisher's z Transformation)									
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		p Value for H0:Rho=0
X	Y	60	-0.86606	-1.31711	-0.00734	-0.86422	-0.916926	-0.781872	<.0001

Through SAS proc corr procedure, we could know the correlation coefficient, r=-0.87 (95%C.I.: -.92, -.78), indicates a strong negative linear association with X and Y.

2) The bivariate analysis shows that there is a significant association between age and muscle mass (p<.0001); with the age increasing, the muscle mass decreased.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	156.34656	5.51226	28.36	<.0001	145.31257	167.38056
X	1	-1.19000	0.09020	-13.19	<.0001	-1.37054	-1.00945

Adv Regression, Homework 1

Yiwen Ren

Beta1-hat= -1.19 | Beta0-hat=156.34656

Got from SAS output Table-Covariance of Estimates:

Var(beta1-hat) = 0.008 | Var(beta0-hat) = 30.385

Got from SAS output Table-Analysis of Variance:

(Sigma-hat)^2 =66.80

(d) Interpretation

$\hat{\beta}_1$: For each year increase in age, the expected muscle mass increase is -1.19.

$\hat{\beta}_0$: when age $=0$, the expected muscle mass is 156.347.

From SAS output above, we could get :

95% CI for $\hat{\beta}_0$ is $(145.313, 167.381)$

95% CI for $\hat{\beta}_1$ is $(-1.371, -1.009)$

(e).

① H_0 : $\beta_1 = 0$ (not linearly associated)

H_1 : $\beta_1 \neq 0$ (linearly associated)

② t-test.

$$T = \frac{\beta_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{(n-1)S_x^2}}} \sim t_{n-2}$$

③ The test statistic's distribution is t-distribution with df = 58 under H_0

④. $T = -13.19$ with d.f. = 58

⑤ Conclusion:

p-value < 0.0001, we reject Null Hypothesis. So, the mean muscle mass is linearly associated with age.

(f). $\bar{x} = 59.98$, the I created $X_i^* = X_i - \bar{X}$ by using SAS. data procedure.

i:

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	84.97063	1.05515	80.53	<.0001	82.85852	87.08275
xnew	1	-1.19000	0.09020	-13.19	<.0001	-1.37054	-1.00945

$\hat{\beta}_0^* = 84.97$ $\quad\quad$ $\hat{\sigma}^{*2} = 66.80$ (Got from ANOVA Table)

$\hat{\beta}_1^* = -1.19$

ii. Interpretation:

β_1^*: For each year increasing in the mean age, the expected muscle mass increase is -1.19.

β_0^*: When the age is equal to mean age, the expected muscle mass is 84.97.

iii: From two SAS output above, we could know the $\beta_1^* = \hat{\beta}_1$ and SE of β_1^* = the SE of $\hat{\beta}_1$, but the SE of β_0^* is far less than the SE of $\hat{\beta}_0$.

iv. the advantage of using X_i^* over X_i:

Using X_i^* gave us a smaller $var(\hat{\beta}_0)$ comparing using X_i directly, which means the model created by the second dataset is more accurate.

(g).

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11627	11627	174.06	<.0001
Error	58	3874.44750	66.80082		
Corrected Total	59	15502			

From the SAS output above, we could get:

$F = 174.06$ with $(1, 58)$ d.f.

Then we could know the straight-line regression of Y on X is significant (p-value < 0.0001).

2.

(a).

$$Y_i = \beta_1 x_i + \varepsilon_i$$

$$SSE = \sum_{i=1}^{n} (Y_i - \beta_1 x_i)^2$$

$$\frac{\partial}{\partial \beta_1} SSE = 0$$

$$= -2 \sum_{i=1}^{n} x_i (Y_i - \beta_1 x_i) = 0$$

Then $\sum_{i=1}^{n} x_i Y_i = \beta_1 \sum_{i=1}^{n} x_i^2$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$$

The estimator $\hat{\beta}_1$ is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$$

Yiwen Ren

(b).

$$mse(\hat{\beta}_1) = E[(\hat{\beta}_1 - \beta_1)^2]$$

$$= Var(\hat{\beta}_1) + bias(\beta_1)^2$$

$\because$ $\hat{\beta}$ is an unbiased estimator of β_1

$\therefore$ $bis(\beta_1)^2 = 0$

$$\therefore \quad mSE(\hat{\beta}_1) = var(\hat{\beta}_1) = \frac{\sigma^2}{SSX}$$

(c) $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{SSX})$

Prove:

$$E(\hat{\beta}) = E(\sum_{i=1}^{n} k_i y_i) = \sum_{i=1}^{n} k_i E(y_i)$$

$$= \sum_{i=1}^{n} k_i (\beta_0 + \beta_1 x_i) = \sum_{i=1}^{n} k_i \beta_0 + \sum_{i=1}^{n} \beta_1 x_i \cdot k_i$$

$$= \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i x_i$$

$$= \beta_0 \cdot 0 + \beta_1 \cdot 1 = \beta_1$$

$$Var(\hat{\beta}_1) = Var(\sum_{i=1}^{n} k_i y_i) = \sum_{i=1}^{n} Var(k_i y_i)$$

$$= \sum_{i=1}^{n} \cdot k_i^2 \cdot Var(y_i)$$

$$= \sigma^2 \cdot \sum_{i=1}^{n} k_i^2$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{SSX}\right)^2 = \frac{\sigma^2}{SSX}$$

$$\therefore \quad \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{SSX})$$

Adv Regression, Homework 1

Yiwen Ren

3. (a).

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
weight	1	0.00527	0.00045806	11.50	<.0001	0.00428	0.00626

From proc reg procedure, $\hat{\beta}_1 = 0.00527$, $\hat{\sigma}^2 = 0.489$

① $H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

② Test Statistics:

$$t = \frac{\beta_1}{\sqrt{\frac{\hat{\sigma}^2}{(n-1)S_x^2}}} = 11.50 \text{ with } df = 12.$$

③ p-value < 0.0001

We reject H_0, accept H_1 that $\beta_1 \neq 0$.

Also, we could see the 95% CI of β_1 doesn't include 0.

(b)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.11598	2.50749	0.05	0.9639	-5.34737	5.57934
weight	1	0.00498	0.00615	0.81	0.4333	-0.00841	0.01837

$\hat{\beta}_0 = 0.11598$

$\hat{\beta}_1 = 0.00498$

Test for $\beta_0 = 0$:

From SAS output above, we could get

$t = 0.05$ with $df = 12$

Then we get p-value $= 0.9639 > 0.05$.

So. we fail to reject H_0, and $\beta_0 = 0$.

Test for $\beta_1 = 0$:

$t = 0.81$ with $df = 12$,

Then we get p-value $= 0.4333 > 0.05$.

So, we fail to reject H_0, and $\beta_1 = 0$.

There is no significant linear relationship between the rat body weights and latency to seizures.

(c). Findings: At (a) part, when we set $\beta_0 = 0$, we got a result that L and W has a linear relationship but when we accounted for β_0 at part (b), we got an. adverse result, indicating that sometimes, the relationship may be spurious. if we failed to account all factors.

Adv Regression, Homework 1

Yiwen Ren

4.

$$t = \frac{N(0,1)}{\sqrt{\chi^2(p)/p}}$$

$$t \sim t_p$$

$$t^2 = \frac{N^2(0,1)}{\chi^2(p)/p} \quad N^2(0,1) \sim \chi^2_{(1)}$$

$$= \frac{\chi^2(1)/1}{\chi^2(p)/p}$$

$$\sim F(1,p)$$

$$t^2 = F(1,p)$$