ST 411/511 Homework 8

Due on March 11

Aashish Adhikari

Winter 2020

Instructions

This assignment is due by 11:59 PM, March 11, 2020 on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile (should you choose – recommended) using the provide .Rmd (R Markdown) template.** If you opt to not use R Markdown, please format your solutions in a similar manner as provided in this document. Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, gramatically correct sentences for your solutions.

Problems (25 points total)

Question 1

(a) (1 point) What is wrong with this formulation of the regression model: $Y = \beta_0 + \beta_1 X$? How would you express it instead?

This formulation means that the response variable, Y has a deterministic value for each value of the explanatory variable X. However, one of the assumptions of the simple (i.e., univariate) linear regression model is that for each value of the explanatory variable X, there exists a distribution of the response variable. Hence, I would instead specify the mean of the response variable as a function of X. i.e., $\mu(Y|X) = \beta_0 + \beta_1 X$

(b) (1 point) What assumptions are made about the distribution of the explanatory variable in the simple linear regression model?

This question got me confused because all the assumptions that I could read from the class slides, the book, and the internet focus in particular to the response variable only. I am listing them out at the end but that would not answer what this question is asking for. I read several articles where the analysts have stated that there are no assumptions about the independent variables in a linear regression model except for that one that they should be independent of each other. They do not have to be normally distributed. But the question asks for siple linear regression with just one explanatory variable. One assumption that holds, however, is that the relationship between the explanatory variable and the response variable should be linear. Also, the observations of the explanatory variable are independent of the observations of the response variable.

Nonetheless, the standard assumptions of simple linear regression the pertain to the explanatory variable in some way are as below: 1.For each value of the explanatory variable, there is a normally-distributed sub-population of responses. 2.The means of the sub-populations fall on a straight ine function of the explanatory variable.

Question 2 (Modified from *Sleuth* 7.27)

Black wheatears, *Oenanthe leucura*, are small birds of Spain and Morocco. Males of the species demonstrate an exaggerated sexual display by carrying heavy stones to nesting cavities. Different males carry somewhat

different sized stones, prompting a study of whether larger stones may signal a higher health status. M. Soler et al. (1999) calculated the average stone mass (grams) carried by each of 21 male black wheatears, along with T-cell response measurements reflecting their immune systems' strengths. The data are in `ex0727`.

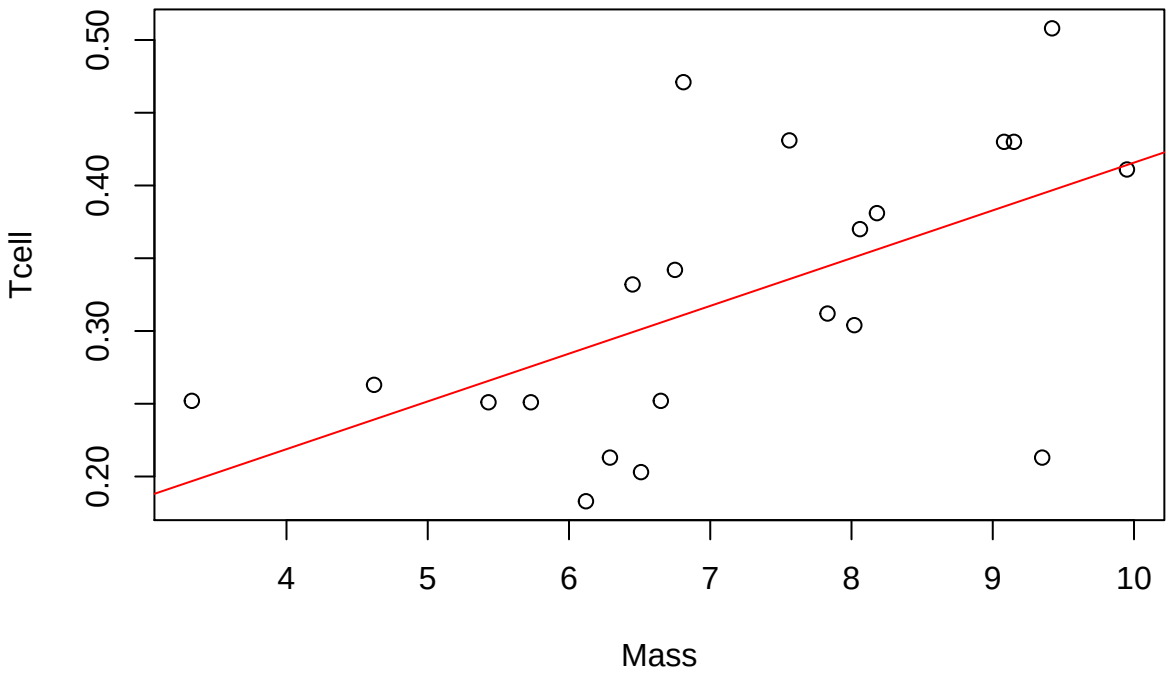
(a) (1 point) Make a scatter plot of `Mass` (X) versus `Tcell` (Y) including the estimated regression line.

```
#get the data set
data(ex0727)
ex0727
```

```
##    Mass Tcell
## 1  3.33 0.252
## 2  4.62 0.263
## 3  5.43 0.251
## 4  5.73 0.251
## 5  6.12 0.183
## 6  6.29 0.213
## 7  6.45 0.332
## 8  6.51 0.203
## 9  6.65 0.252
## 10 6.75 0.342
## 11 6.81 0.471
## 12 7.56 0.431
## 13 7.83 0.312
## 14 8.02 0.304
## 15 8.06 0.370
## 16 8.18 0.381
## 17 9.08 0.430
## 18 9.15 0.430
## 19 9.35 0.213
## 20 9.42 0.508
## 21 9.95 0.411
```

```
#Plot the scatterplot
plot(ex0727$Mass, ex0727$Tcell, xlab = "Mass", ylab = "Tcell")

#Add the regression line, first argument is the response variable
abline(lm(ex0727$Tcell~ex0727$Mass), col="red")
```



(b) (2 points) Fit the linear model using the `lm()` function to regress `Tcell` on `Mass` (i.e., model the mean of `Tcell` as a function of `Mass`). Use the `summary()` function to view more information about the estimated regression model. Provide an interpretation for the p-values of the regression coefficients.

```
lin_reg_model <- lm(ex0727$Tcell~ex0727$Mass)
summary(lin_reg_model)
```

```
##
## Call:
## lm(formula = ex0727$Tcell ~ ex0727$Mass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18138 -0.04673  0.01796  0.04219  0.15999
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08750    0.07868   1.112  0.27996
## ex0727$Mass  0.03282    0.01064   3.084  0.00611 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08102 on 19 degrees of freedom
## Multiple R-squared:  0.3336, Adjusted R-squared:  0.2986
## F-statistic: 9.513 on 1 and 19 DF,  p-value: 0.006105
```

The p values can be seen above. Interpretation: (Applies to both the coefficients) –> Assuming that the true coefficient is 0, the probability that we will observe a t-statistic greater or more extreme than the absolute value of the observed t-statistic is 0.27996 (for β_0 and 0.00611 for β_1)

For β_0, it is greater than the significance level and thus we cannot reject the hypothesis that y-intercept is 0.

For β_1, it is less than the significance level and thus we reject the hypothesis that the slope is 0.

(c) (1 point) Construct 90% confidence intervals for the regression parameters using the `confint()` function.

```
#Fit the model first
model1 <-lm(Tcell~Mass, data = ex0727)
confint(model1, level = 0.9)
```

```
##                      5 %       95 %
## (Intercept) -0.04854518 0.22353914
## Mass         0.01442095 0.05122203
```

The confidence intervals for both the coefficients can be found above.

(d) (1 point) Estimate the mean T-cell measurement for a new bird that is observed to carry stones averaging 4 grams in weight by using the `predict()` function. Construct a 95% confidence interval for *mean* T-cell measurement for that new bird.

```
predict(model1, newdata=data.frame("Mass"=4), interval="confidence", level=0.95)
```

```
##         fit       lwr       upr
## 1 0.2187829 0.1383913 0.2991746
```

The estimate for the mean T-cell measurement for a bird that is observed to carry stones averaging 4 grams in weight is 0.218. The confidence interval lower bound is 0.138 and the upper bound is 0.299.

(e) (1 point) Construct a 95% *prediction* interval for T-cell measurement for the new bird in part (d). How does the prediction interval compare to the confidence interval from part (d)?

```
predict(model1, newdata=data.frame("Mass"=4), interval="prediction", level=0.95)
```

```
##         fit        lwr       upr
## 1 0.2187829 0.03111638 0.4064495
```

The prediction interval lower bound is 0.031 and the upper bound is 0.406. Note that the prediction interval is wider than the confidence interval. I believe it is because in the prediction interval, we are predicting the plausible values for any possible T-cell measurement while in the confidence interval, we are bounding the mean value for the given mass.

Question 3

Suppose that the estimated simple linear regression of a response Y on a predictor X based on $n = 6$ observations produces the following residuals:

```
resid <- c(-0.09, 0.18, -0.27, 0.16, -0.06, 0.09)
```

Note: For this question, all of the computations should be performed "by-hand".

(a) (1 point) What is the estimate of σ^2?

```
residual_square = resid^2
sample_sigma = sum(residual_square)/(6-2)
sample_sigma
```

```
## [1] 0.037675
```

The estimate of variance is 0.037.

(b) (2 points) Further, you know that the estimated regression parameters are $\hat{\beta}_0 = -0.54$ and $\hat{\beta}_1 = 0.08$. Additionally, the sample mean of X is 13.5 and the sample variance of X is 15.5. Find the standard errors of the two estimated regression parameters.

The standard errors of β_0 and β_1 can be obtained by replacing the true population standard deviation with an estimate in the formulas for their standard deviations. (From Book 7.3.5)

Given, sample $\beta_0 = 0.54$ sample $\beta_1 = 0.08$ sample mean of X's = 13.5 Sample Variance of X's = 15.5

```
n <- 6
sample_var <- 15.5
sample_mean <- 13.5
beta_0=-0.54
beta_1=0.08

SE_of_beta_0 <- sqrt(sample_sigma)*sqrt(((1/n)+(sample_mean^2/((n-1)*sample_var))))
SE_of_beta_1 <- sqrt(sample_sigma*(1/((n-1)*sample_var)))

SE_of_beta_0
```

```
## [1] 0.3080198
```

```
SE_of_beta_1
```

```
## [1] 0.02204833
```

The standard error of β_0 is 0.3080198. The standard error of β_1 is 0.02204833.

(c) (2 points) Use the standard error you calculated in (b) to test the null hypothesis $H_0 : \beta_1 = 0$ that the true population slope has value 0. State your test statistic, two-sided p-value, and what you conclude from the test.

```
sample_beta_0 <- -0.54
sample_beta_1 <- 0.08
t_statistic <-(sample_beta_1 - 0)/SE_of_beta_1
t_statistic
```

```
## [1] 3.628392
```

```
beta_1_p_val <- 2 *  (1- pt(abs(t_statistic),df = 4))
beta_1_p_val
```

```
## [1] 0.02219148
```

The t-statistic for $H_0 : \beta_1 = 0$ is 3.628392. The p-value is 0.02219. We reject the null hypothesis that the slope i.e., β_1 could be zero.

(d) (1 point) What is the mean value of Y we would predict when $X = 12$? (E.g., what is $\hat{\mu}(Y|X = 12)$?)

```
mean_y <- sample_beta_0+(sample_beta_1*12)
mean_y
```

```
## [1] 0.42
```

The mean value for X = 12 is 0.42.

(e) (2 points) Calculate the standard error of $\hat{\mu}(Y|X = 12)$ and use this value, along with your result from part (d), to find a 95% confidence interval for mean Y when $X = 12$.

```
X_0 <- 12
standard_error_hat=sqrt(sample_sigma*((1/n)+((X_0-sample_mean)^2/((n-1)*sample_var))))
standard_error_hat
```

```
## [1] 0.08586592
```

```
conf_int_lower_bound=mean_y-(qt(0.975,df=4)*standard_error_hat)
conf_int_upper_bound=mean_y+(qt(0.975,df=4)*standard_error_hat)

conf_int_lower_bound
```

```
## [1] 0.181598
```

```
conf_int_upper_bound
```

```
## [1] 0.658402
```

The standard error is 0.08586. The required confidence interval ranges from 0.1815 to 0.658.

(f) (2 points) Calculate the standard error of $Pred(Y|X = 12)$ and use this value, along with your result from part (d), to find a 95% prediction interval for mean Y when $X = 12$.

```
se_pred_hat=sqrt(sample_sigma*(1+(1/n)+((X_0-13.5)^2/((n-1)*sample_var))))
se_pred_hat
```

```
## [1] 0.212245
```

```
pred_int_upr=mean_y+(qt(0.975,df=4)*se_pred_hat)
pred_int_lw=mean_y-(qt(0.975,df=4)*se_pred_hat)

pred_int_lw
```

```
## [1] -0.1692867
```

```
pred_int_upr
```

```
## [1] 1.009287
```

The standard error is 0.212245. The prediction interval upper and lower limits are -0.1692867 and 1.009287 respectively for the mean when X=12.

(g) (1 point) The sample correlation between X and Y is 0.8831. Find the value of R^2 for the regression considered in the previous parts of this question.

```
corre <-0.8831
rsq <- corre^2
rsq
```

```
## [1] 0.7798656
```

The r^2 value is 0.7798656.

Question 4 (Modified from *Sleuth* 8.26)

The ex0826 data set contains the average mass, average metabolic rate, and average lifespan for 95 species of mammals. Kleiber's law states that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power 0.75. Judge the adequacy of this theory with these data by following these steps:

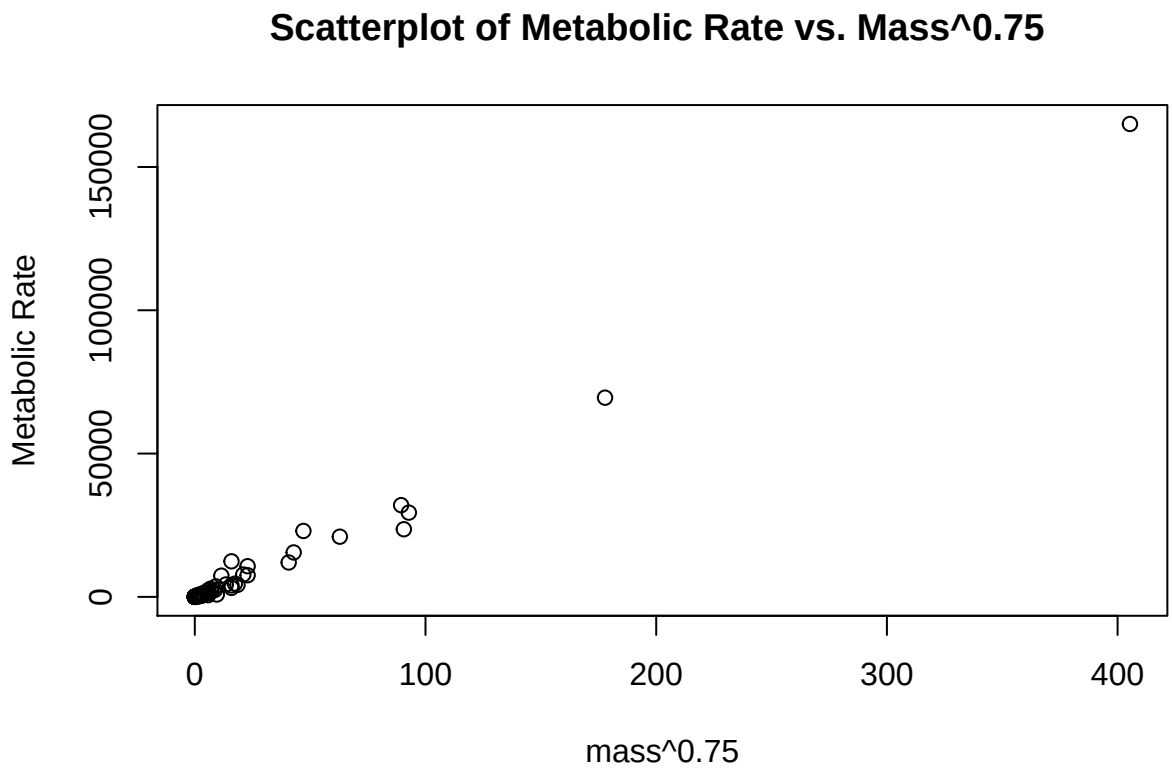
(a) (1 point) Make a scatterplot of metabolic rate (Y) versus mass$^{0.75}$ (X) for these 95 species.

```
names(ex0826)
```

```
## [1] "CommonName" "Species"    "Mass"       "Metab"      "Life"
```

```
mass <- ex0826$Mass
mass <- mass^0.75
metabolic <- ex0826$Metab

plot(metabolic~mass,data = ex0826,xlab="mass^0.75",ylab = "Metabolic Rate",main="Scatterplot of Metaboli
```

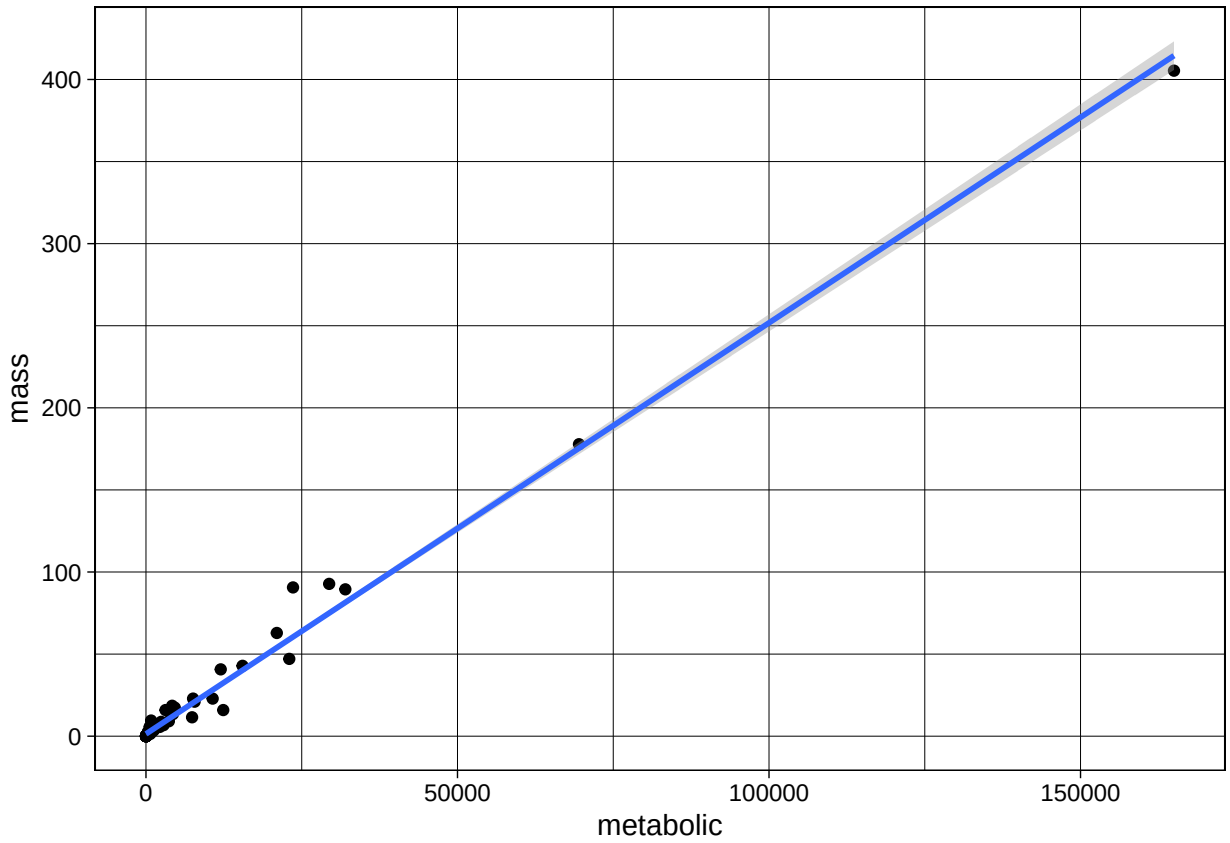


(b) (1 point) Fit a linear regression model of metabolic rate (Y) regressed on mass$^{0.75}$ (X). Provide the estimated coefficients, estimated standard deviation $\hat{\sigma}$, and R^2. (You need to indicate what these are in the R output – don't just include the R output.)

```
model2 <- lm(metabolic~mass,data=ex0826)
summary(model2)
```

```
##
## Call:
## lm(formula = metabolic ~ mass, data = ex0826)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11712.7   -117.4    368.5    474.3   6598.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -481.346    213.467  -2.255   0.0265 *
## mass         395.016      4.299  91.895   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1992 on 93 degrees of freedom
## Multiple R-squared:  0.9891, Adjusted R-squared:  0.989
## F-statistic:  8445 on 1 and 93 DF,  p-value: < 2.2e-16
```

```
ggplot(data=ex0826, aes(metabolic,mass))+
  geom_point()+
  theme_linedraw()+
  geom_smooth(method = "lm",se=T)
```



Estimated coefficients are beta_0_hat : -481.346 and beta_1_hat : 395.016.

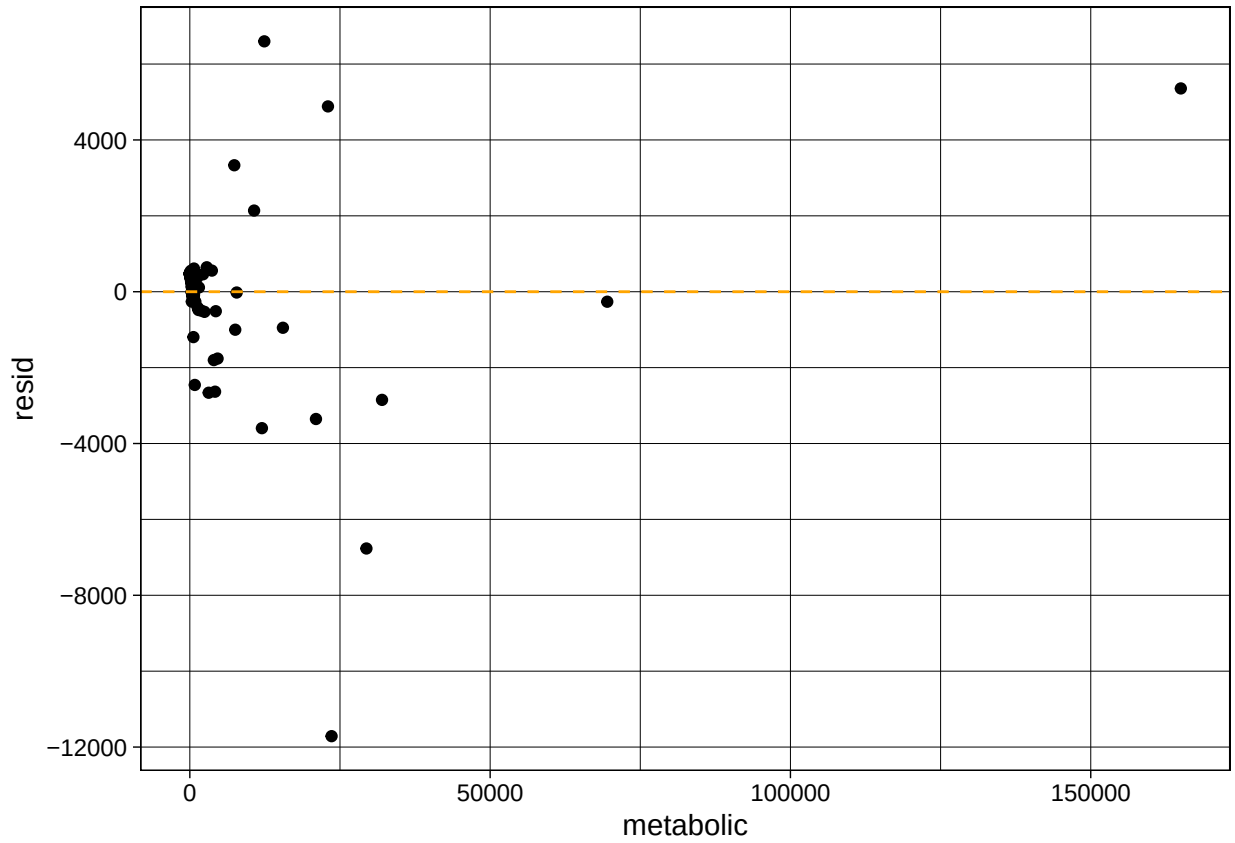
The standard deviation is 1992.

The R^2 values are a)Multiple Rsquared: 0.9891 b) Adjusted Rsquared: 0.989.

(c) (2 points) Plot the residuals vs. the fitted values from the model fit in part b). Examine the plot and discuss whether i) the linear fit seems appropriate; and ii) the assumptions for simple linear regression inference to be valid appear to be met.

```
ex0826$resid <- model2$resid

ggplot(data = ex0826, aes(metabolic, resid)) +
  geom_point() +
  theme_linedraw() +
  geom_hline(yintercept = 0, lty = 2, color="orange")
```



It looks like as the values of metabolic rate varies on x direction, there is an equal variance in the residuals along the y axis. Thus, it satisfies the following four assumptions:

1. The mean of each response is related to the explanatory variable by a linear function: $\mu(Y|X) = \beta_0 + \beta_1 X$.
2. All pairs of points, (explanatory, response) are statistically independent.
3. The variance of each response is the same for all values of the explanatory variable.(Same spread)
4. All responses are taken from normal distributions. (and are also independent as we know that mass, species, metaboli rate,etc are independent)

(d) (1 points) It has also been suggested that metabolic rate is one of the best single predictors of species lifespan. Fit a linear regression model of lifespan (Y) regressed on metabolic rate (X). Provide the estimated coefficients, estimated standard deviation $\hat{\sigma}$, and R^2. (You need to indicate what these are in the R output – don't just include the R output.)

```
model3 <- lm(ex0826$Life~metabolic, data=ex0826)
summary(model3)
```

```
##
## Call:
## lm(formula = ex0826$Life ~ metabolic, data = ex0826)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.205  -6.885  -2.341   3.598  61.775
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 1.030e+01  1.124e+00   9.161 1.22e-14 ***
## metabolic   3.873e-04  5.740e-05   6.748 1.26e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.57 on 93 degrees of freedom
## Multiple R-squared:  0.3287, Adjusted R-squared:  0.3215
## F-statistic: 45.53 on 1 and 93 DF,  p-value: 1.262e-09
```

The estimated coefficients are $\beta_0 = 1.030\text{e+}01$ and $\beta_1 = 3.873\text{e-}04$. The estimated standard deviation (error) is 10.57. The multiple R^2 is 0.3287 and the adjusted R^2 is 0.3215..

(e) (1 point) How much variation in the distribution of mammal lifespans can be explained by metabolic rate?

The extent of variation in the distribution of mammal lifespans that can be explained by the metabolic rate is given by multiple R-squared value 0.3287.