ST 411/511 Homework 6

Pavan Sai Nallagoni

Summer 2022

Instructions

This assignment is due by 11:59 PM, August 12th on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile using the provide .Rmd (R Markdown) template.**

> Note: Create a PDF by either compiling a PDF directly (via LaTeX) or from a Word Doc. Do not submit a PDF of the HTML output as they're cumbersome and difficult to read.

Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope (Failure to do so will result in point deductions). You should also write complete, grammatically correct sentences for your solutions.

Once you've completed the assignment, and before you submit it to Gradescope, you should read the document to ensure that (1) The computed values show up in the document, (2) the document "looks nice" (i.e. doesn't have extraneous code/outputs and includes *just* the essentials), and (3) ensure the document is "easy" to read.

Goals:

1. Understand the theoretical structure and ideas behind simple linear regression models.
2. Practice making scatterplots and linear regression models in R.
3. Practice interpreting R's linear regression outputs.
4. Learn how to assess the uncertainty in estimates for regression coefficients and model predictions through the use of confidence and prediction intervals.
5. Learn the basics of variable transformations in linear regressions and assessing the assumptions of linear regression through diagnostic plots.

Question 1 (2 points)

(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?

In Regression setting, we model the average value of Y for a given value of X as a linear function of X and not the individual values and can be is represented as:

$$\mu(Y|X) = \beta_0 + \beta_1 X$$

Another way of implementing is to consider an error term which is represented as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

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

Usually, we don't make any explicit assumptions about the distribution of the explanatory variable in the simple linear regression model. However, if they're present in our samples then there's a correlation such that they'll have their impact on the residuals.

One particular assumption about the explanatory variable in the simple linear regression model is that they are independent to give the best fit.

Question 2 (6 points) - Modified from *Sleuth* 7.27

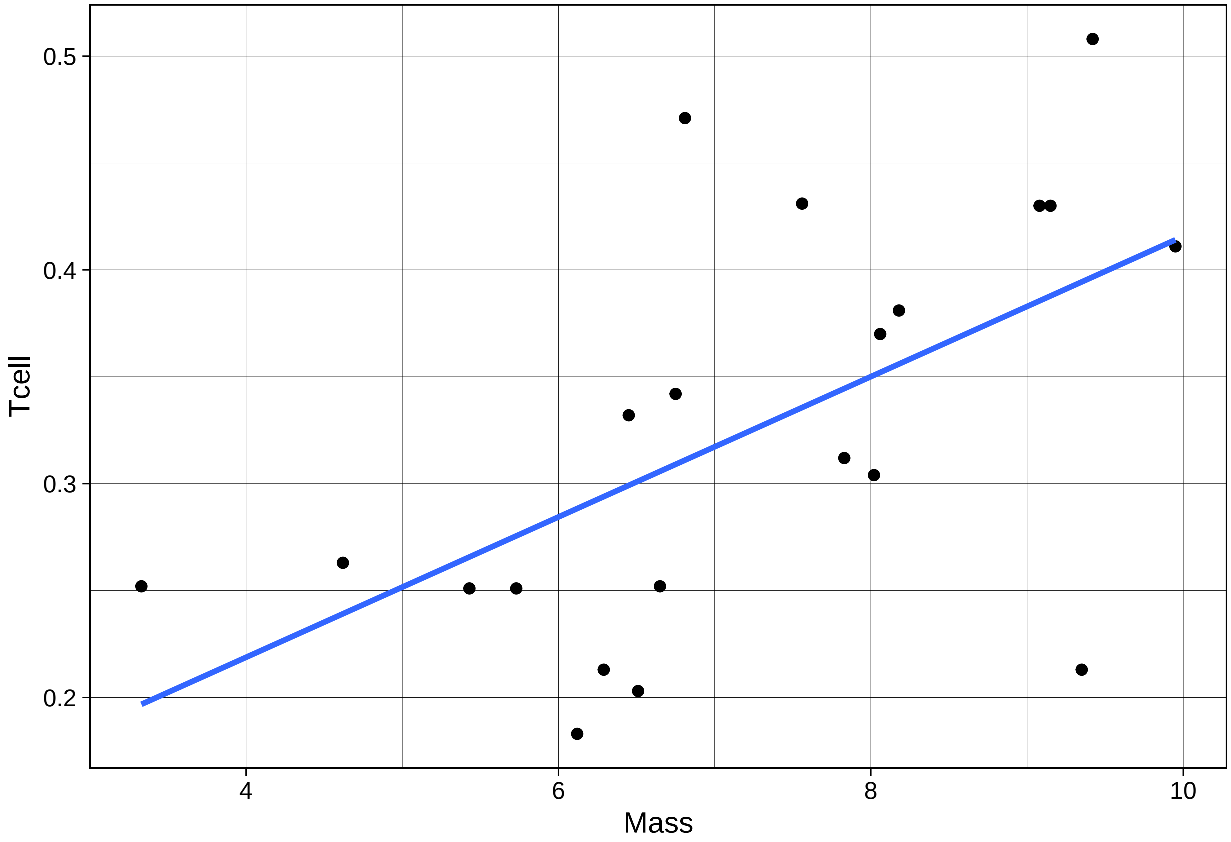
Black wheatears, *Oenanthe leucura*, are small birds of Spain and Morocco. Males of the species demonstrate an exaggerated sexual display by carrying heavy stones to nesting cavities. Different males carry somewhat different sized stones, prompting a study of whether larger stones may signal a higher 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.

```
data(ex0727)
names(ex0727)
```

```
## [1] "Mass"  "Tcell"
```

```
ggplot(data = ex0727, aes(Mass, Tcell)) +
  geom_point() +
  theme_linedraw() +
  geom_smooth(method="lm", se=FALSE)
```



(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 regression coefficients and state whether or not they are statistically significant and why.

```
mod1 <- lm(Tcell ~ Mass, data=ex0727)
summary(mod1)
```

```
##
## Call:
## lm(formula = Tcell ~ Mass, data = ex0727)
##
## 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
## 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
```

p-values for $\beta_0 = 0.27996$

p-values for $\beta_1 = 0.00611$

We fail reject the null hypothesis that the intercept β_0 is 0 since the p-value $= 0.27996$ is greater than the significance level $\alpha = 0.05$, and conclude that we have no evidence against the null hypothesis that the true T-cell count is zero when the stone mass carried is zero.

We reject the null hypothesis that the slope β_1 is 0 since the p-value $= 0.00611$ is less than the significance level $\alpha = 0.05$, and conclude that we do have evidence for a linear trend between Mass and mean Tcell.

(c) (1 point) Construct 90% confidence intervals for the regression parameters using the `confint()` function. Interpret the confidence interval you construct for the slope parameter.

```
confint(mod1, level=0.90)
```

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

The 90% confidence interval for the slope parameter is (0.01442095, 0.05122203). Therefore, we can say that, with 90% confidence, the true slope parameter describing how Tcell changes with a one unit increase in Mass lies in the range 0.01442095 to 0.05122203.

(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. Interpret your 95% confidence interval.

```r
predict(mod1, newdata=data.frame("Mass"=4))
```

```
##         1
## 0.2187829
```

```r
predict(mod1, newdata=data.frame("Mass"=4), interval="confidence", level=0.95)
```

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

The 95% confidence interval of the mean of Tcell lies in the interval (0.1383913, 0.2991746) given Mass = 4. Therefore, with 95% confidence, we can estimate that the mean T-cell measurement for a new bird that is observed to carry stones averaging 4 grams in weight lies in the range 0.1383913 and 0.2991746.

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

```r
predict(mod1, newdata=data.frame("Mass"=4), interval="prediction", level=0.95)
```

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

The 95% prediction interval of the mean of Tcell lies in the interval (0.03111638, 0.4064495) given Mass = 4.Therefore, we can estimate that an individual observation will likely be predicted to be in the interval (0.03111638, 0.4064495).

Question 3 (11 points)

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". Make sure that the outputs from the code that end up in your document are *just* the requested value(s) and not any intermediate steps.

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

```
n <- 6 # Given
df <- n - 2
sigma2 <- sum(resid^2)/(df)
sigma2
```

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

(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.

```
sigma <- sqrt(sigma2)
sub_problem_b0 <- sqrt(1/n + 13.5^2 / ((n-1) * 15.5))

se_b0 <- sigma * sub_problem_b0
se_b0
```

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

```
sub_problem_b1 <- sqrt(1/((n-1)*15.5))
se_b1 <- sigma * sub_problem_b1
se_b1
```

```
## [1] 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.

```
b1 <- 0.08
t <- b1/se_b1
t
```

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

```
p <- 2*(1-pt(abs(t), df))
p
```

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

We would reject the null hypothesis $H_0 : \beta_1 = 0$ that the true population slope has value 0 since the p-value is less than the significance level $\alpha = 0.05$. Hence, there's is linearity in the mean since Y is valid and a function of X.

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

```
b0 <- -0.54

Y <- b0 + b1*12
Y
```

```
## [1] 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 $X = 12$.

```
sub_problem <- sqrt(1/n + (12-13.5)^2 / ((n-1) * 15.5))
se_sub <- sigma*sub_problem
se_sub
```

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

```
lower_bound <- Y - qt(0.975, df)*se_sub
lower_bound
```

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

```
upper_bound <- Y + qt(0.975, df)*se_sub
upper_bound
```

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

(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 $X = 12$.

```
sub_p <- sqrt(1+1/n+(12-13.5)^2/((n-1)*15.5))

se_p <- sigma * sub_p
se_p
```

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

```
plower_bound <- Y - qt(0.975, df)*se_p
plower_bound
```

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

```
pupper_bound <- Y + qt(0.975, df)*se_p
pupper_bound
```

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

(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.

```
rXY<- 0.8831

R2 <- (rXY)^2
R2
```

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

Question 4 (8 points) - Modified from *Sleuth* 8.26

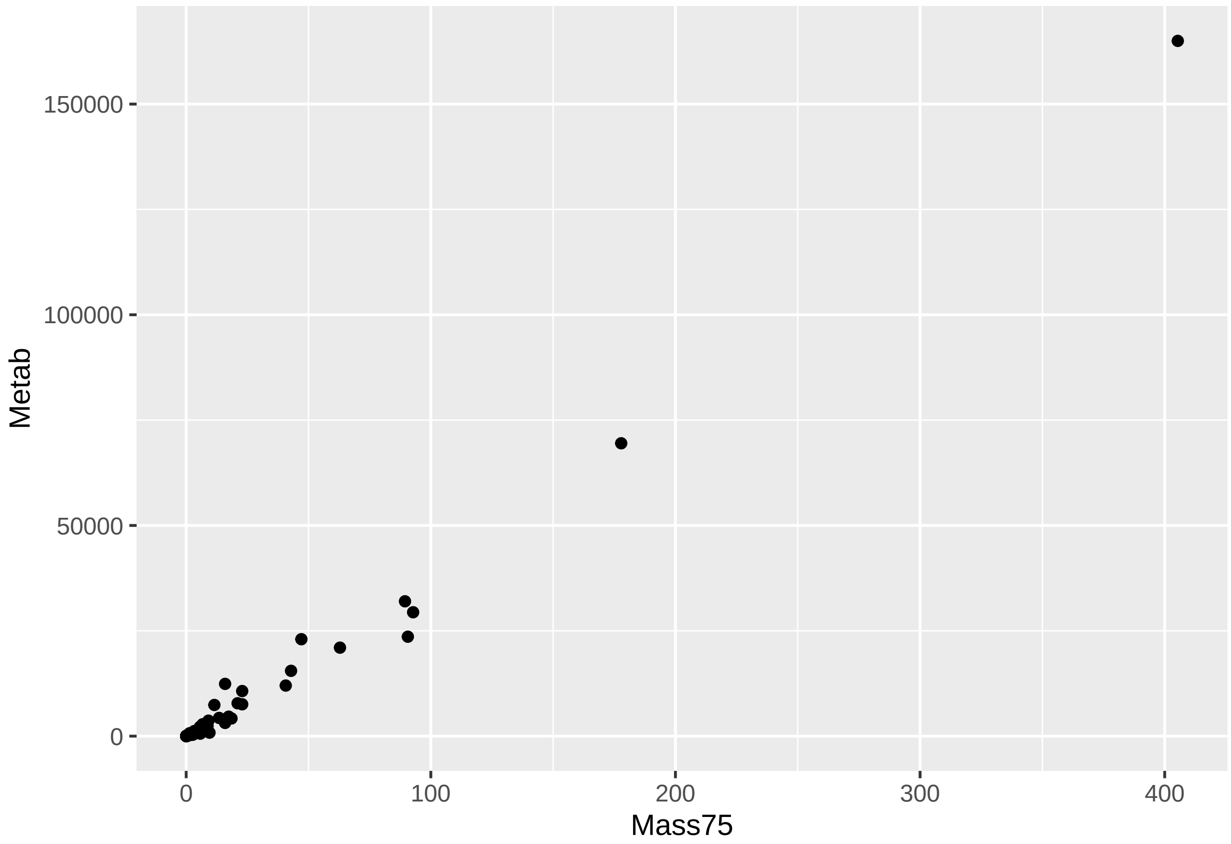
The ex0826 data set contains the average mass, average metabolic rate, and average lifespan for 95 species of mammals. Kleiber's law states that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power 0.75. Judge the adequacy of this theory with these data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

(a) (1 point) Make a scatterplot of metabolic rate (Y) versus mass$^{0.75}$ (X) for these 95 species. Based on your scatterplot, does fitting a linear model seem like an appropriate method for describing the relationship between metabolic rate and mass$^{0.75}$?

```
data(ex0826)
names(ex0826)
```

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

```
ex0826$Mass75 <- ex0826$Mass^0.75
ggplot(data = ex0826, aes(Mass75, Metab)) +
  geom_point()
```



Based on your scatterplot, fitting a linear model seem like an appropriate method for describing the relationship between metabolic rate and mass$^{0.75}$ as all the points are close to each other and may lie on the line.

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

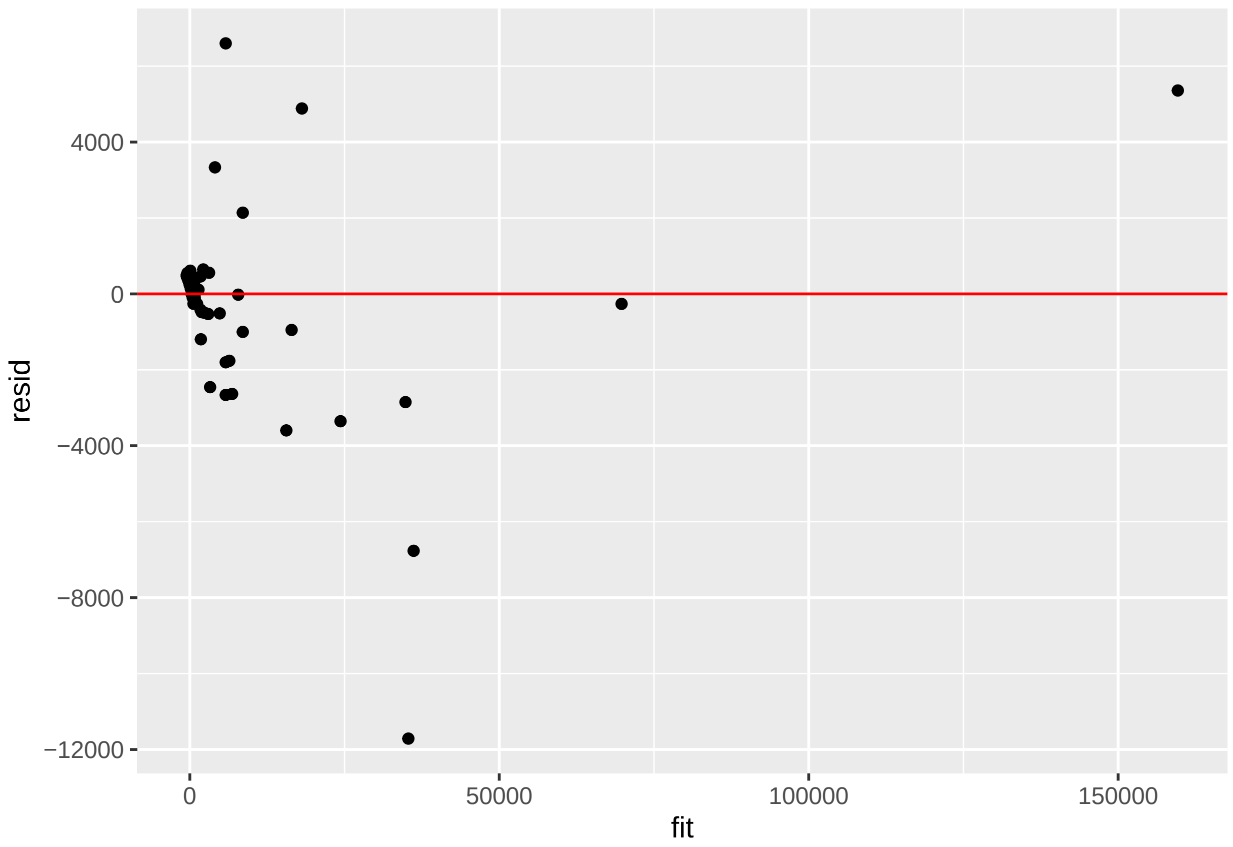
```
mod4 <- lm(Metab ~ Mass75, data=ex0826)
summary(mod4)
```

```
##
## Call:
## lm(formula = Metab ~ Mass75, 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 *
## Mass75       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
```

- Estimated intercept coefficient: -481.346
- Estimated slope coefficient: 395.016
- Estimated standard error (sigma-hat): 1992
- Estimated R-squared: 0.9891

(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$fit <- mod4$fit
ex0826$resid <- mod4$resid
ggplot(ex0826, aes(x=fit, y=resid)) +
  geom_point() +
  geom_hline(yintercept=0, color="red")
```

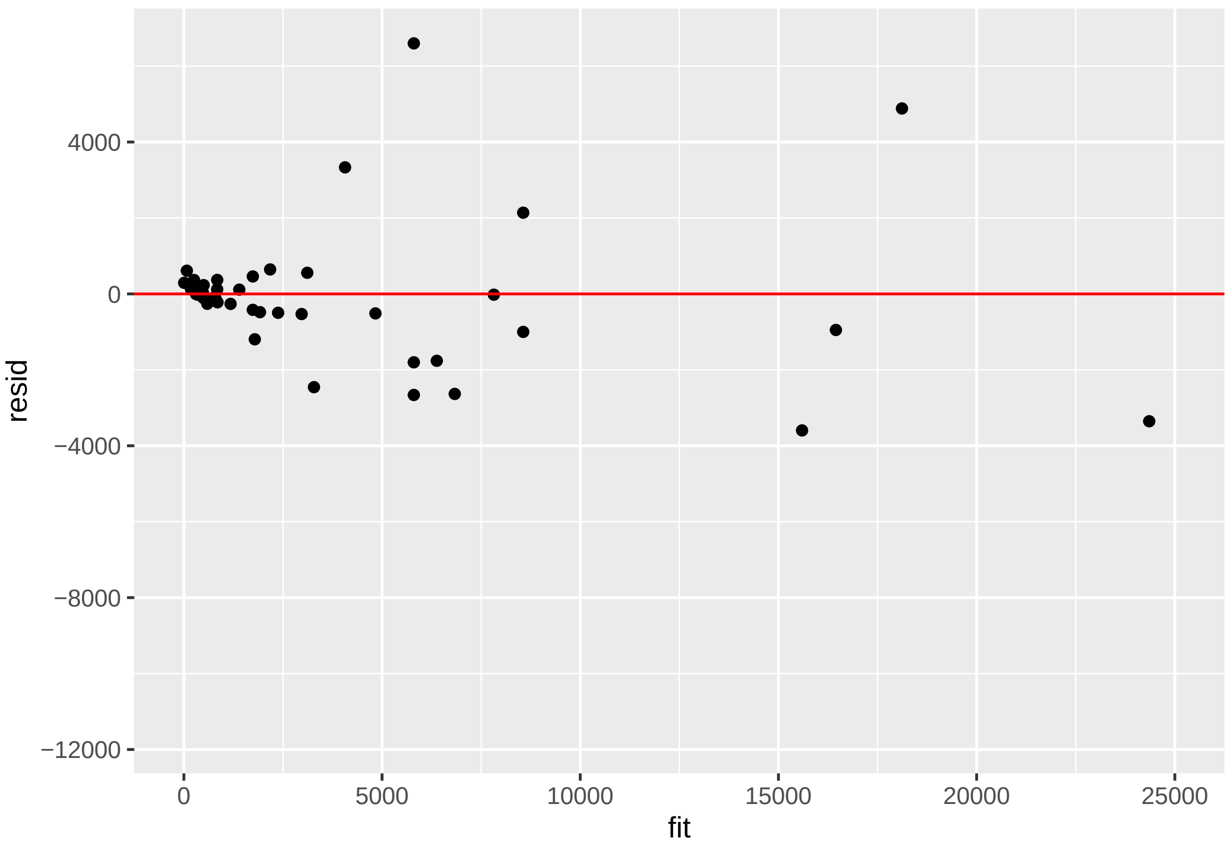


Form the above plot, the linear fit seems appropriate since most of the values are closer to each other. However, we can't say they have a constant variance since the points doesn't form a clear 'U' shape.

(d) (2 points) Plot the residuals vs. the fitted values from the model fit in part b) again, but this time, adjust the limits of the x-axis to view just the data points whose fitted values are between 0 and 25,000 (You can consider setting the x-axis values to 10,000 as well to really "Zoom in" on the residuals). Does looking at this "zoomed-in" diagnostic plot change your assessment of whether or not the linear regression assumptions are met? Why?

> Note: You can update the limits of your plot, assuming you're using `ggplot2`, by "adding the layer": `scale_x_continuous(limits=c(0, 25000))`

```
ggplot(ex0826, aes(x=fit, y=resid)) +
  geom_point() +
  geom_hline(yintercept=0, color="red") +
  scale_x_continuous(limits=c(0, 25000))
```



It is clear that linearity assumption is validated here as well. However, the points are spread and are far from the each other results in a non-constant variance and it might violate the homoskedasticity rule. Finally, since the sample is not big enough we can't say much about the normality, and the pair of observations remain independent.

(e) (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 error $\hat{\sigma}$, and R^2. (You need to indicate what these are in the R output – don't just include the R output.)

```
mod4d <- lm(Life ~ Metab, data=ex0826)
summary(mod4d)
```

```
##
## Call:
## lm(formula = Life ~ Metab, 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 ***
## Metab       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
```

- Estimated intercept coefficient: 1.030e+01
- Estimated slope coefficient: 3.873e-04
- Estimated standard error (sigma-hat): 10.57
- Estimated R-squared: 0.3287

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

The percent of variation in the distribution of mammal lifespans that can be explained by metabolic rate is given by the R^2 value. The Estimated R^2 value is 0.3287 which means that 32.87% variation in the distribution of mammal lifespans that can be explained by metabolic rate.