1. (20 pts) Describe (in words, and mathematical formulae where appropriate) the meanings of the following concepts:

(a) Residual:

Difference b/w what we observed and what our model predicts (fitted value) ~~our model predicts (fitted value) and what we actually observed~~

obs - expected

$y_i - \hat{y}_i$ for some x_i

good!

(b) Coefficient of Determination (R^2)

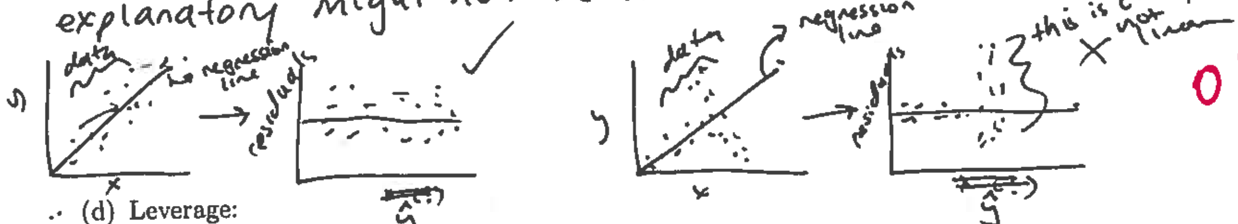
r^2 How much variance is explained by the model ÷ ~~vs.~~ how much variance there is total — or, how well the regression line fits the data ✓

$r^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$

or, does our model tell us anything that the mean doesn't?

(c) Heteroskedasticity (non-constant variance): of residuals?

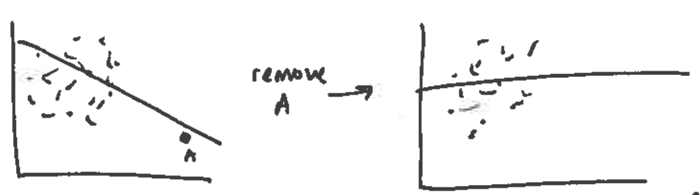
You want the residuals to have constant variance ~~$\varepsilon \sim N(0, \sigma_\varepsilon)$~~ no, that is normally distributed w/ linear regression b/c if they don't, it means that ~~a linear~~ the relationship b/w your response & explanatory might not be linear ✓



OK

(d) Leverage:

whether removing an influential outlier has a lot of effect on the regression line:

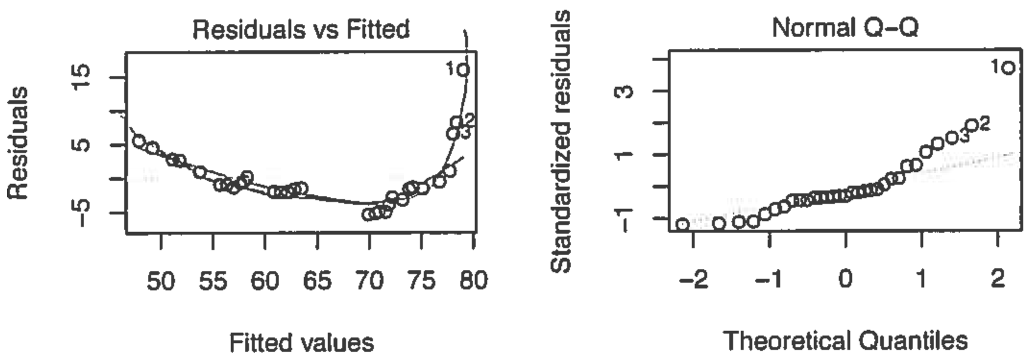


thus, A had high leverage b/c removing it drastically changed the slope of

good!

2. (15 pts) Refer to the simple linear regression model below for the women's world record *time* in the 100m freestyle swimming event as a function of the *year* in which the record was set.

```
mod = lm(time ~ year, data=filter(SwimRecords, sex=="F"))
par(mfrow=c(1,2))
plot(mod, which=c(1,2))
```



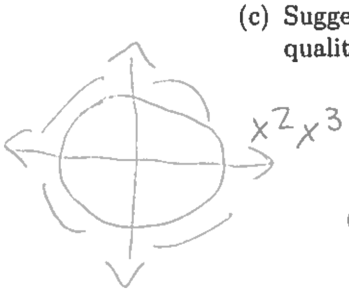
(a) Is the condition for **Linearity** met? Why or why not?

No, there is a clear U pattern in the Residuals vs. Fits graph. If linearity was met this plot would show a straight rather than curved line.

(b) Is the condition for **Normality** met? Why or why not?

The Normal Q-Q plot shows the standardized residuals vs. theoretical perfectly normal ones. If the condition of normality was met the points would follow the dotted line. in the plot above it variates from the line in the upper right corner. The condition of Normality is not met.

(c) Suggest a possible transformation for the explanatory variable that might improve the quality of the fit. Why do you think this transformation would fit the data better?



Using Tukey and the Residuals vs. Fits plot I would suggest squaring the explanatory variable. The data clearly bulges to the right in the U shape in the plot, and using the Tukey graphic I would square the explanatory variable. I would also get rid of point 1 as it lies above 3 on standardized Residuals making it a clear outlie

good

(a) Interpret the meaning of the slope coefficient in the context of the real-world problem. **Be sure to include units!**

For every inch taller the father is we expect an averge increase in height by 0.43 inches for children of the same sex.

(b) Interpret the meaning of the coefficient for *sex* in the context of the real-world problem. **Be sure to include units!**

We expect male children to be an average of 5.18 inches taller then female children with fathers of the same height

(c) Alex was a man (in 19th century Britain) whose father was 72 inches tall. What height does the model predict for Alex?

$$\widehat{Alex} = 34.46 + 0.43(72) + 5.18(1)$$
$$= 70.6 \text{ inches}$$

(d) Is a parallel slopes model appropriate in this context? Why or why not?

Yes because sex is a categorical variable. The scatterplot also supports this because the lines on the plot were not made using the parallel slopes method instead the sexes were moded using SLR separately however their slopes are still nearly parallel.

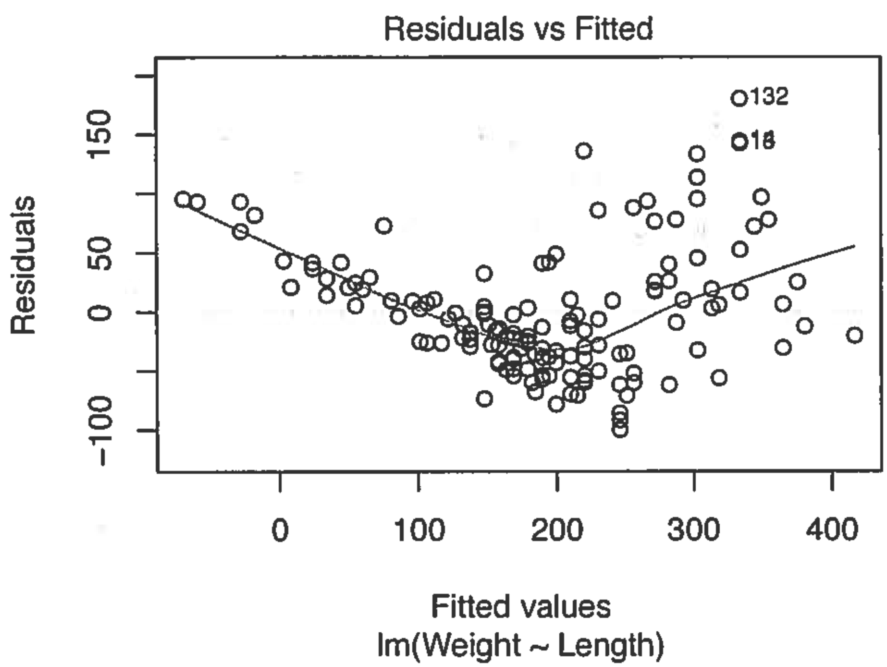
good!

9

4. (8 pts + 3 pts) The **Bears** data set contains data on 143 young bears. Consider the following simple linear regression model for the *Weight* of a bear (in pounds) as a function of its *Length* (in inches).

y x

```
Bears = read.csv("http://www.math.smith.edu/~bbaumer/mth247/labs/Bears.csv")
mod = lm(Weight ~ Length, data=Bears)
plot(mod, which=1)
```



(a) Is the condition for **Constant Variance** met? Why or why not?

No. The variability of the errors is not the same ✓
There is a certain pattern
And also when the length goes higher up, the residuals become much bigger ✓
It's a fan shaped pattern

(b) (Extra credit, 3 pts) Suggest, with justification, a possible transformation of the response variable that might improve the situation.

+3

We could carry out a square root or log transformation of the response variable year. We could tell the fan shape "<" on the right, where the bears of higher length tend to have a wider range of errors. We need to transform in order to shrink the values of the weight and the larger values needed to be affected more than smaller ones ✓

12

5. (12 pts) Consider a model for the fuel economy of several cars as a function of the size of the car's engine. The response variable is *hwy* (measured in miles per gallon) and the explanatory variable is *displ* (the volume of the engine, measured in liters).

```
mod = lm(hwy ~ displ, data=EPA)
coef(mod)

## (Intercept)       displ
##   33.311977   -3.140088
```

$\hat{hwy} = 33.31 - 3.14(\hat{displ})$

(a) Suppose you are considering the purchase of a new car, and you are debating between a Nissan 350Z, which has a 3.5 liter engine, and a Mazda MX-5, which has a 2.0 liter engine. Which car would the model predict to get better gas mileage? Why?

Mazda MX-5 because the displ. variable has a negative effect on the mileage. Therefore, a smaller engine has a greater fuel economy

(b) What is the maximum gas mileage that the model could predict for a given car? Interpret this number in a real-world context. Is it meaningful?

According to the model, the max mileage is 33.311977. This is meaningless because it corresponds to an engine of ØL, which is impossible.

(c) Suppose now that you are considering the purchase of a specific car with a 4.3 liter engine. Write down an interval that you are 95% sure contains the fuel economy of that car.

```
newcars = data.frame("displ" = c(4.3))
predict(mod, newdata=newcars, interval="confidence")

##       fit      lwr      upr
## 1 19.8096 19.48721 20.13198

predict(mod, newdata=newcars, interval="predict")

##       fit      lwr      upr
## 1 19.8096 12.80012 26.81908
```

We are 95% confident that this car has a gas mileage between 12.80012 and 26.81908.

15

6. (15 pts) The following ANOVA table refers to the regression model for fuel economy (*hwy*) as a function of engine size (*displ*) on the previous page.

```
anova(mod)

## Analysis of Variance Table
##
## Response: hwy
##             Df Sum Sq Mean Sq F value    Pr(>F)
## displ        1  12549 12549.1   986.1 < 2.2e-16 ***
## Residuals 834  10614    12.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) What percentage of the variation in fuel economy (*hwy*) is explained by the model?

$$r^2 = 1 - \frac{10614}{(12549+10614)} = 0.54$$ ✓

54% of variation is explained by the model

(b) What the value of the correlation coefficient between *hwy* and *displ*?

Slope of SLR is negative, therefore r should be negative

$r = -\sqrt{r^2} = -0.74$ ✓

NICE!

(c) How confident are you that our model provides statistically significant information about the fuel economy (*hwy*) of cars? Justify your answer.

$F = 986.1 > 1$ and P-value < 0.001 which is less than our α level $\approx 5\%$ therefore we can reject H_0 hypothesis and conclude that our model provides statistically significant info about the fuel economy of cars ✓

9

7. (9 pts) The data set **KidsFeet** contains data on the *length* and *width* of the feet of several children. Use the statistics below to answer the following questions.

```
favstats(~length, data=KidsFeet)

##  min Q1 median   Q3  max     mean       sd  n missing
## 21.6 24   24.5 25.6 27.5 24.72308 1.317586 39       0

favstats(~width, data=KidsFeet)

## min   Q1 median   Q3 max     mean        sd  n missing
## 7.9 8.65      9 9.35 9.8 8.992308 0.5095843 39       0

cor(length ~ width, data=KidsFeet)

## [1] 0.6410961
```

$\widehat{\text{length}} = \hat{\beta}_0 + \hat{\beta}_1 * \text{width}$

(a) What is the value of the slope coefficient for a simple linear regression model for *length* as a function of *width*?

$\hat{\beta}_1 = r_{xy} \cdot \frac{Sd\,y}{Sd\,x} = 0.641096 \cdot \frac{1.317586}{0.5095843} = 1.6576241$

The slope $\hat{\beta}_1 = 1.6576241$

(b) What is the value of the corresponding intercept?

$\hat{\beta}_0 = \hat{Y} - \hat{\beta}_1 \overline{X} = 24.72308 - 1.6576241 * 8.992308$
$= 9.8172130$

The intercept $\hat{\beta}_0 = 9.8172130$

(c) What is the value of the coefficient of determination (R^2) for that model?

$R^2 = (0.641061)^2$
$\simeq 0.411$

The coefficient of determination $R^2 = 41.1\%$