Multiple Regression

Linear Regression: Model Diagnostics

- Conditions for least squares (linear) regression

    1. **Linearity.** The data should show a linear trend (linear relationship between x and y).
        - Check condition using residual plot (see end of Homework 8 for example of how to make dataframe to plot residuals using ggplot)
    2. **Nearly normal residuals.** Residuals must be nearly normal
        - Check condition using Q-Q plot or histogram of residuals (see Ch.4 R lab for a refresher)
    3. **Constant variability.** The spread of points around the least squares line is roughly constant.
        - Check condition using residual plot
    4. **Independent observations.** Observations should be independent of each other (this is all we will work with in this class). Examples of when this is violated include time series data, e.g., daily stock price or daily temperature - in these cases there is an underlying structure to how these observations are related that has to be taken into account.
        - Check condition using residual plot

    *** See pg. 319, Figure 8.12 for nice examples of when these conditions are violated.

Example - Regression with two predictors

Let's consider the data set from Homework 8, which we used to look at the relationship between mother's smoking status and gestational age. We saw (or will see) that smoking status alone doesn't have the expected effect on gestational age, and the literature tells us that we need to consider a combination of smoking status and mother's age to explain more of the variability we see in gestational age. This means we want a linear model with more than one explanatory variable. Let's try two: smoking status and mother's age.

```r
babies <- read.csv("https://www.openintro.org/stat/data/csv/babies.csv")
```

Model fit and inference

```r
## Include only gestation and smoke varibles; omit all NA values
babies3 <- na.omit(babies[,names(babies) %in% c("gestation", "smoke", "age")])

## Multiple regression model with gestation as response, smoke and age as predictors:
lm_fit2 <- lm(gestation ~ smoke + age, data=babies3)
summary(lm_fit2)
```

```
##
## Call:
## lm(formula = gestation ~ smoke + age, data = babies3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -132.097   -6.934    0.726    8.372   72.549
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 285.0492     2.2785 125.102   <2e-16 ***
## smoke        -2.3266     0.9434  -2.466   0.0138 *  
## age          -0.1769     0.0799  -2.213   0.0271 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 16.02 on 1208 degrees of freedom
## Multiple R-squared:  0.008454,   Adjusted R-squared:  0.006812 
## F-statistic:  5.15 on 2 and 1208 DF,  p-value: 0.00593
```

Looking at the summary for the model fit, we see that we have:

$$\hat{y} = 285.05 - 2.33(smoke) - 0.18(age)$$

where smoke is 1 if the mother smokes, and 0 otherwise, and age is the mother's age.

We can test for a relationship between mother's age and gestational age:

$H_0 : \beta_{age} = 0$ versus $\beta_{age} \neq 0$

We calculate the test statistic, $t = -2.213$ and the p-value, 0.0271. There is moderate evidence that age is negatively associated with maternal age. This means that on average, as maternal age increases, gestational age decreases (which isn't desirable).

We can test to see if smoking status is related to gestational age. If it isn't, it shouldn't matter if the mother smokes or not, i.e.,

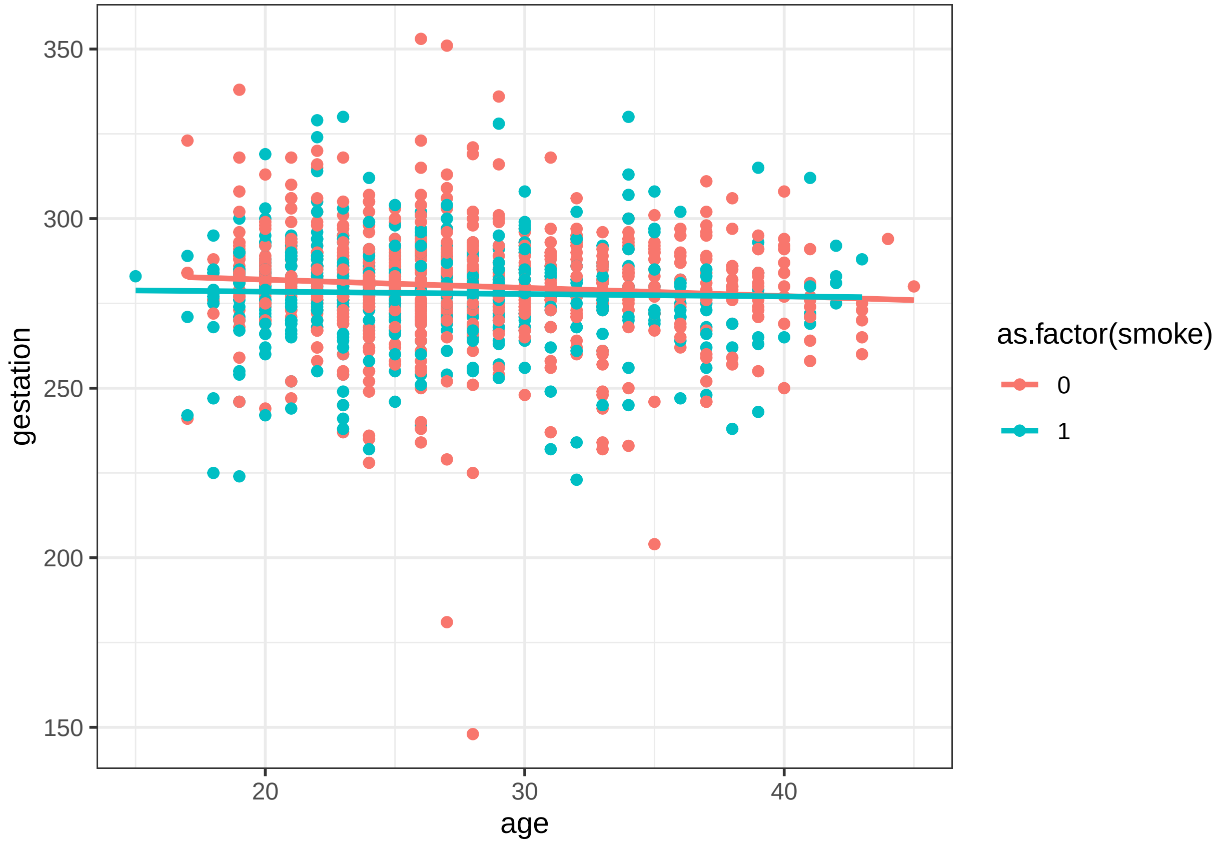
$H_0 : \beta_{smoke} = 0$ versus $\beta_{smoke} \neq 0$

We calculate the test statistic, $t = -2.433$ and the p-value, 0.0138. There is moderate evidence that smoking status is associated with maternal age. If a mother smokes, the gestational age decreases (the point estimate for mean gestational age for a smoking mother is 2.33 days less than that for a non smoking mother of the same age).

Check linear model conditions (linearity, independence, equal variance, normal residuals)

Linearity:

```r
## Scatterplot to visualize this relationship:
ggplot(data=babies3, aes(x=age, y=gestation, color=as.factor(smoke))) +
  geom_point() +
  geom_smooth(method="lm", se = FALSE) +
  theme_bw()
```



```
## There may be two different lines for the groups. Note, you will see more of this if you continue on
## There are also potential outliers around age = 27, 28, 35; STAT 242 teaches formal methods for ident
```

There isn't an obvious non-linear pattern present, but this also doesn't look like a strong linear relationship. In other words, it's not obvious that we should expect any other relationship (other than linear), but we shouldn't be surprised that our most recent model only accounts for 0.8% of the variability in gestational age. Biological processes can be complicated!

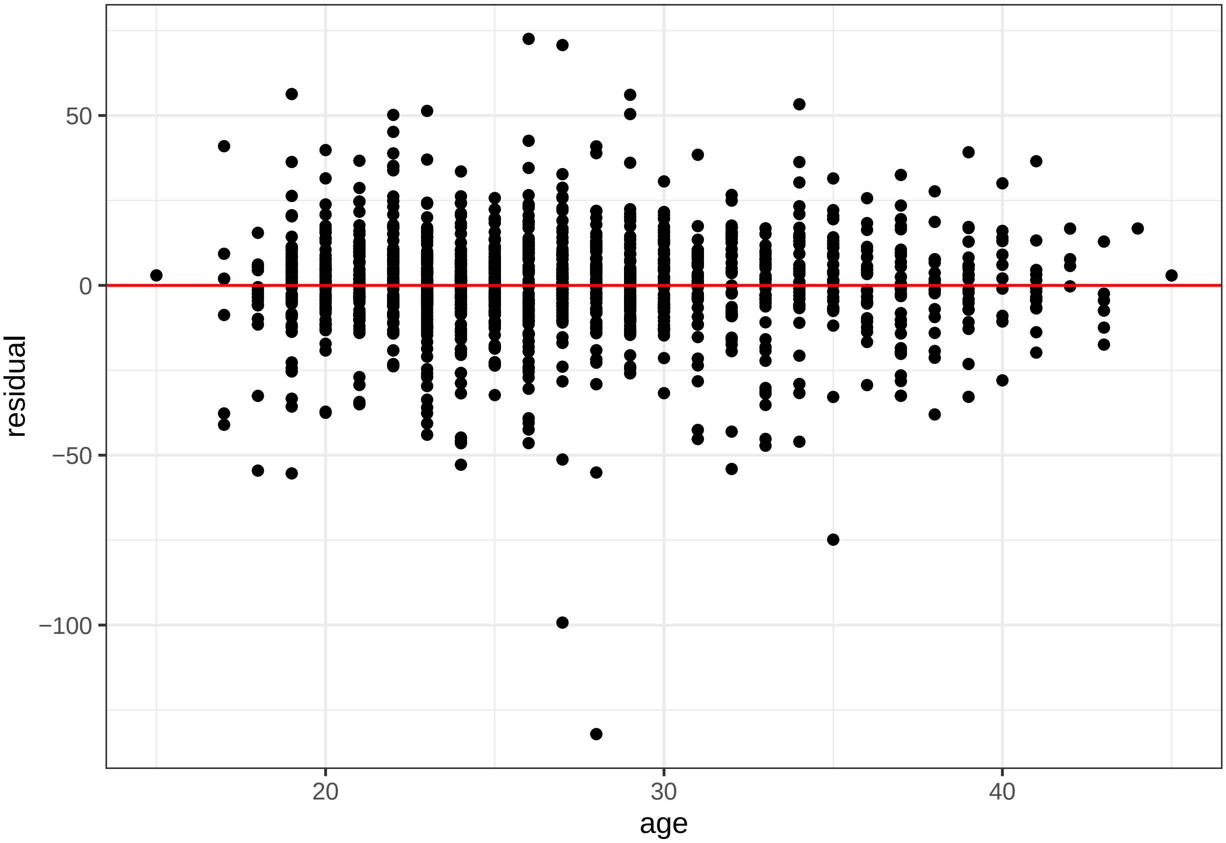
Independence:

This depends on how the data were collected. Check data sources to see if this is a random sample.

Same variance:

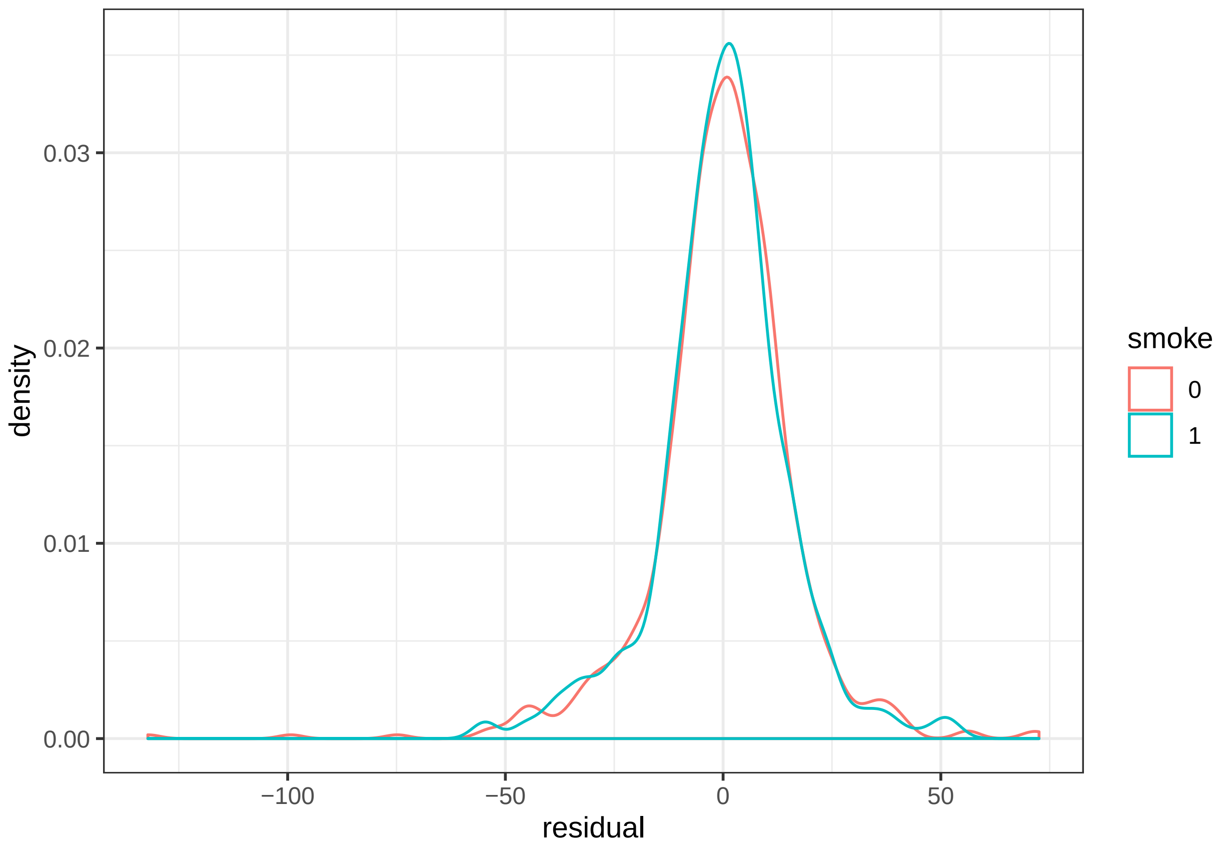
```
## Residual plot:
residuals.df <- data.frame(age=babies3$age,
                           smoke=as.factor(babies3$smoke),
                           residual=residuals(lm_fit2))

ggplot(data=residuals.df, aes(x=age, y=residual)) +
  geom_point() +
  geom_hline(yintercept=0, color="red") +
  theme_bw()
```



Apart from the outliers, this looks good. Recall we are checking to see if there are any obvious patterns in the residual plot (left over after we accounted for the predictor variables).

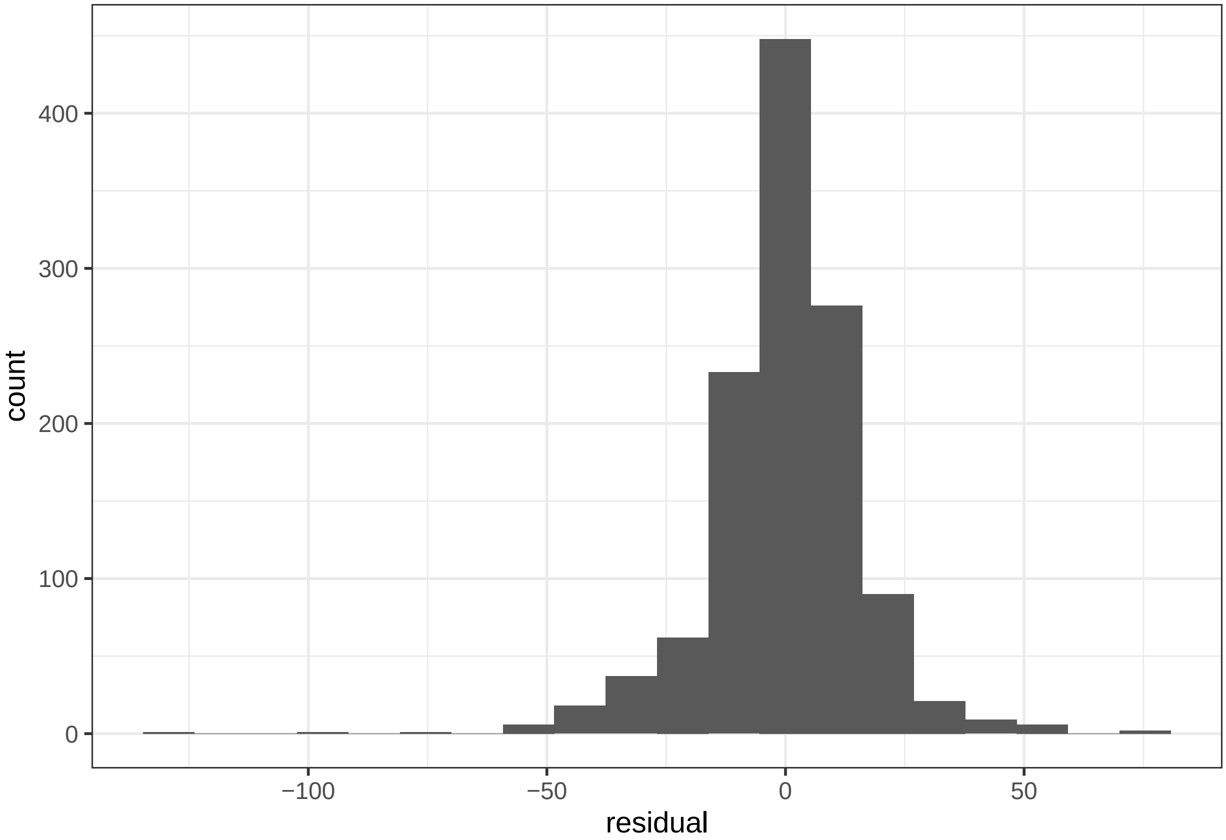
```
## Residual density plot (for categorical predictors):
ggplot(data=residuals.df, aes(x=residual, color=smoke)) +
  geom_density() +
  theme_bw()
```



Here we are looking for these to be unimodal and really similar - this looks good; equal variance, and normality for the residuals are satisfied as a function of smoking status.

Nearly normal residuals:

```
ggplot(data=residuals.df, aes(x=residual)) +
  geom_histogram(bins=20) +
  theme_bw()
```



This isn't awful, although it is left-skewed (potential outliers are present). We might consider a transformation (see intermediate statistics class).

Lab 9, Part II Assignment

The objective of this part of the assignment is to give you a little practice with multiple regression. You will get a chance to work with the more (and get feedback on your work) with a different data set for the final project in this class.

In your R Markdown file from this week, do the following using the SAT data set:

1. Fit a new linear model that uses both HSGPA and SATSum as explanatory variables. Name this linear model lm_var2.
2. Print the summary of lm_var2.
3. Interpret the coefficient for HSGPA. In your interpretation, be sure to hold SATSum constant (so you can discuss only the mean change in FYGPA as a result of a one unit change in HSGPA).
4. Interpret the coefficient for SATSum. In your interpretation, be sure to hold HSPGA constant (so you can discuss only the mean change in FYGPA as a result of a one unit change in SATSum).
5. Check the four conditions for the model. You will need to make 5 plots (two scatterplots, two residual plots, and one histogram), as well as comment on independence. Discuss whether the conditions seem satisfied.
6. Compare the R-squared value for lm_var2 to that which you obtained for your simple linear regression model from last time. Interpret your R-squared value for your lm_var2 model.