3.36pt

Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

September 9, 2017

The Voinovich School of Leadership and Public Affairs

Table of Contents

3.36pt

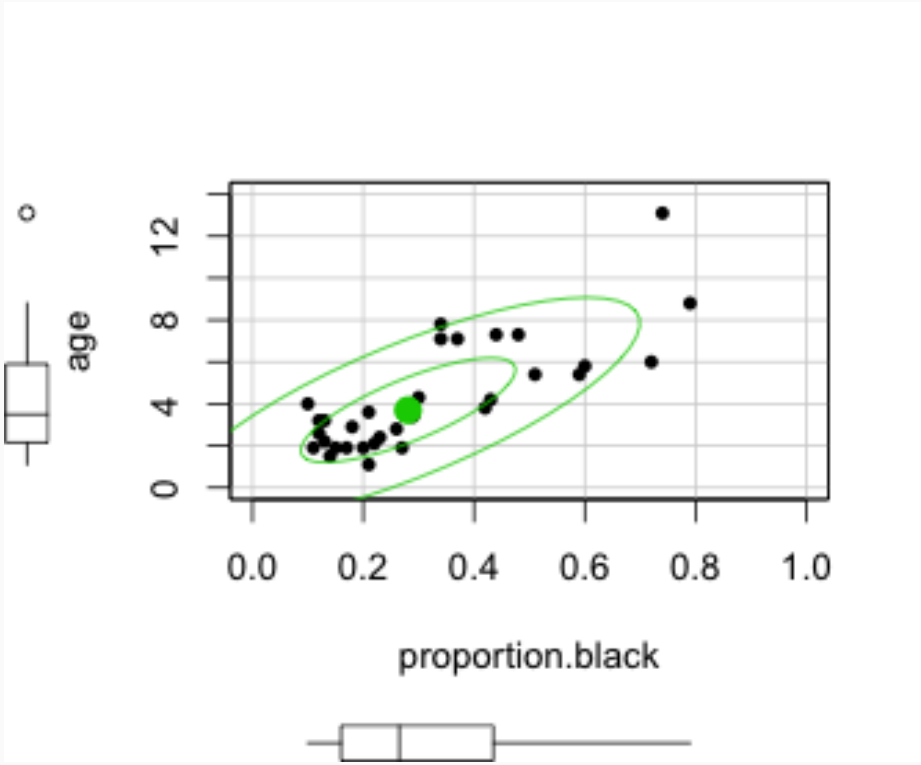
Simple Linear Regression

Introduction to Regression Analysis

- Regression analysis (a) describes and (b) predicts relationships between one continuous or categorical dependent variable and one or more continuous and/or categorical independent variables
- The relationship between y and x is assumed to be linear such that a straight line $y = a + b(x)$ best fits the joint distribution of (x, y)
- Recall the equation for a straight line $y = mx + c$ where $c =$ the intercept, and $m =$ the slope of the line
- In the regression setting
  - a is the intercept (i.e., the value of y when $x = 0$), and
  - b is the slope coefficient
- The slope coefficient (b) tells us how much does y change when x increases or decreases by a unit amount

The Lion's Nose

Lion populations can be controlled by many means but trophy hunting is one way to do it. Knowing the lion's age helps because removing males older than six years of age has little impact on the pride's social structure but killing younger males is more disruptive. Researchers have shown that the amount of black pigmentation on a lion's nose increases with age and so can be used to estimate wild lions' ages. The relationship between age and the proportion of black pigmentation on 32 male lions with known ages is shown below.



Linear Regression with LionNoses

```
> lm1 <- lm(age ~ proportion.black, data=LionNoses)
> summary(lm1)

Call:
lm(formula = age ~ proportion.black, data = LionNoses)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5449 -1.1117 -0.5285 0.9635 4.3421

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.8790    0.5688   1.545    0.133
proportion.black 10.6471   1.5095   7.053 7.68e-08 ***
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1        1

Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared: 0.6238,  Adjusted R-squared: 0.6113
F-statistic: 49.75 on 1 and 30 DF, p-value: 7.677e-08
```

Thus $y = 0.8790 + 10.6471(proportion.black)$

When $proportion.black = 0.20$ predicted $y = 0.8790 + 10.6471(0.20) = 3.00842$

When $proportion.black = 0.21$ predicted $y = 0.8790 + 10.6471(0.21) = 3.114891$

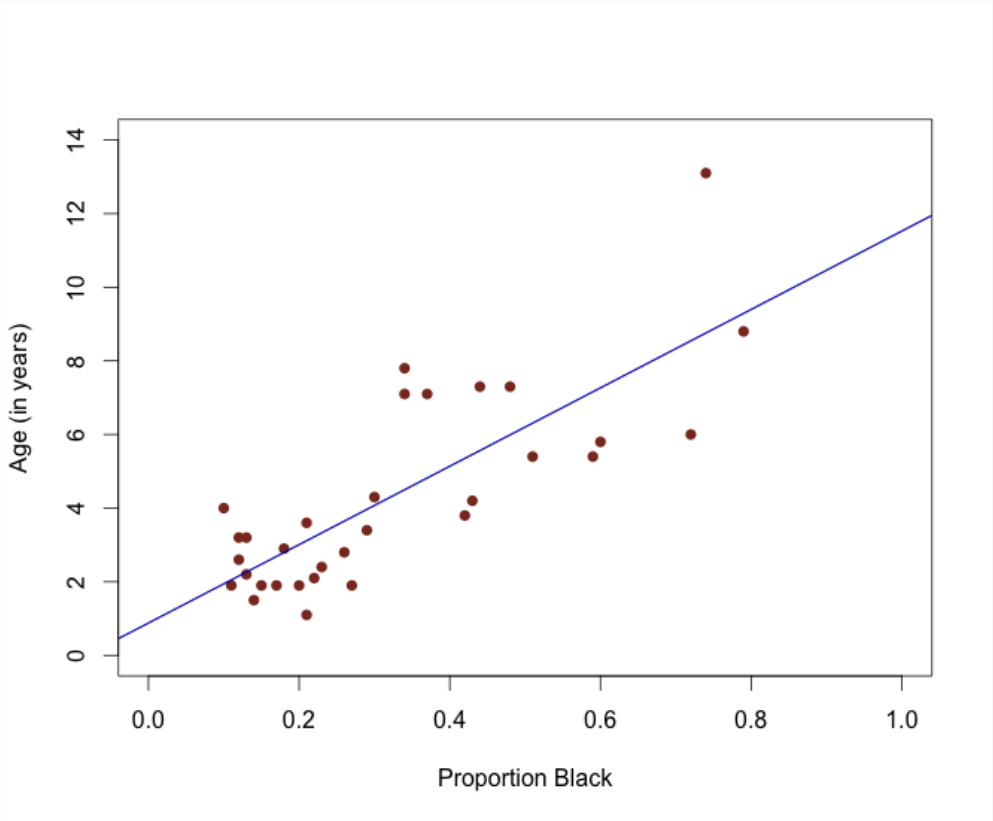
... as proportion.black increased by 0.01 we expect y to increase by 0.106471

In the dataset we actually have lions with 0.20 and 0.21 proportion of their noses black. How old were these lions? The former was 1.9 years old and the latter was 3.6

So the regression equation is making a prediction error because predicted ages were 3.00 and 3.11, respectively!

Unfortunately, with real-world data, you will always have prediction errors; how large or small these will be depends upon how closely and linearly related are x and y, and the quality of your sample

These errors are basically the difference between actual y values and predicted $\hat{y}$ values ... $e = (y - \hat{y})$



The Method of Ordinary Least Squares

OLS looks to minimize $\sum(e_i)^2 = \sum(y_i - \hat{y}_i)^2$

But what is $\sum(y_i - \hat{y}_i)^2$? The Sum of Squared Errors (i.e., SSE)

The estimated intercept and slope are denoted by a ^ symbol and the estimated regression equation is itself written as $\hat{y} = \hat{a} + \hat{b}(x)$

Intercept and the slope are estimated as follows: $\hat{b} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ where $\bar{x}$ is the sample mean of x and $\bar{y}$ is the sample mean of y, the numerator is the covariance of x and y, and the denominator is the Sum of Squares of x

Once we have $\hat{b}$ we can calculate $\hat{a}$ via $\bar{y} = \hat{a} + \hat{b}(\bar{x})$, i..e, $\hat{a} = \bar{y} - \hat{b}(\bar{x})$

```
> lm1 <- lm(age ~ proportion.black, data=LionNoses)
> summary(lm1)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.8790    0.5688   1.545    0.133
proportion.black 10.6471   1.5095   7.053 7.68e-08 ***
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1        1
Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared: 0.6238,  Adjusted R-squared: 0.6113
F-statistic: 49.75 on 1 and 30 DF, p-value: 7.677e-08
```

Breaking Apart the Analysis

- A perfect fit would occur if every y_i were predicted perfectly
- But this rarely occurs. Instead, some or all y_i will $\neq \hat{y}_i$
- $\hat{e}_i = y_i - \hat{y}_i$ is thus called the **residual**
- Summing the squares of all prediction errors yields ...
  the Sum of Squares due to Error $(SS_{residual}) = \sum(y_i - \hat{y}_i)^2$
- What if we calculate $y_i - \bar{y}$ for all i?
- Then we have the Sum of Squares Total $(SS_{total}) = \sum(y_i - \bar{y})^2$
- Sum of Squares due to Regression $(SS_{regression}) = \sum(\hat{y}_i - \bar{y})^2$
- $SS_{total} = SS_{regression} + SS_{residual}$
- Perfect fit occurs when $SS_{residual} = 0$, and thus $SS_{total} = SS_{regression}$
- Abysmal fit occurs when $SS_{regression} = 0$, and thus $SS_{total} = SS_{residual}$
- $R^2 = \frac{SS_{regression}}{SS_{total}}$ thus yields a measure of the "goodness of fit"
  1. $0 \leq R^2 \leq 1$
  2. $R^2 \to 1$ indicates better fit
  3. $R^2 \to 0$ indicates poorer fit

Calculating other elements of the regression equation

- Let us calculate the variance of the residuals $Var(e_i) = \frac{\sum(e_i - \bar{e})^2}{n-2}$
- We know, however, that $\bar{e} = 0$
- Therefore, $Var(e_i) = \frac{\sum(e_i)^2}{n-2} = \frac{\sum(y_i - \hat{y})^2}{n-2} = \frac{SS_{residuals}}{n-2} = MS_{residual}$
- But this is Mean Squared Error (i.e., prediction errors in squared units)
- So if we take $\sqrt{MS_{residual}}$ we get average prediction errors
- Now, the standard error of $\hat{b} = s.e._{(\hat{b})} = \sqrt{\frac{MS_{residual}}{\sum(x_i - \bar{x})^2}}$
- Is this estimate of b significant?
  Proportion black has no impact on age (i.e., $H_0 : \beta_0 = 0$)
  Proportion black has an impact on age (i.e., $H_A : \beta_0 \neq 0$)
- The test statistic is $t_{\hat{b}} = \frac{\hat{b} - \beta_0}{s.e._{(\hat{b})}} = \frac{\hat{b} - 0}{s.e._{(\hat{b})}} = \frac{\hat{b}}{s.e._{(\hat{b})}}$
- We can also test $H_0 : a = 0$; $H_1 : a \neq 0$ via $t_{\hat{a}} = \frac{\hat{a}}{s.e._{(\hat{a})}}$ but this is usually of little substantive interest

Identifying the Elements in R

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.8790    0.5688   1.545    0.133
proportion.black 10.6471   1.5095   7.053 7.68e-08 ***
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1        1
Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared: 0.6238,  Adjusted R-squared: 0.6113
F-statistic: 49.75 on 1 and 30 DF, p-value: 7.677e-08
```

The Estimate of the (Intercept) is $\hat{a} = 0.8790$ and the Estimate of the slope of proportion.black is $\hat{b} = 10.6471$

The standard errors are given for both $\hat{a}$ and $\hat{b}$, and so also the test statistic for each (i.e., the t value)

$P - value$ is also listed for $\hat{a}$ and $\hat{b}$ but as $Pr(> |t|)$ and with symbols ... $*$ means the $P - value < 0.05$; $**$ means the $P - value < 0.01$; $***$ means the $P - value < 0.001$

$R^2 = \frac{SS_{regression}}{SS_{total}}$ is listed as the Multiple R-squared

Adjusted R-squared $= 1 - (1 - R^2)\frac{n-1}{n-k-1}$ where k is no. of independent variables

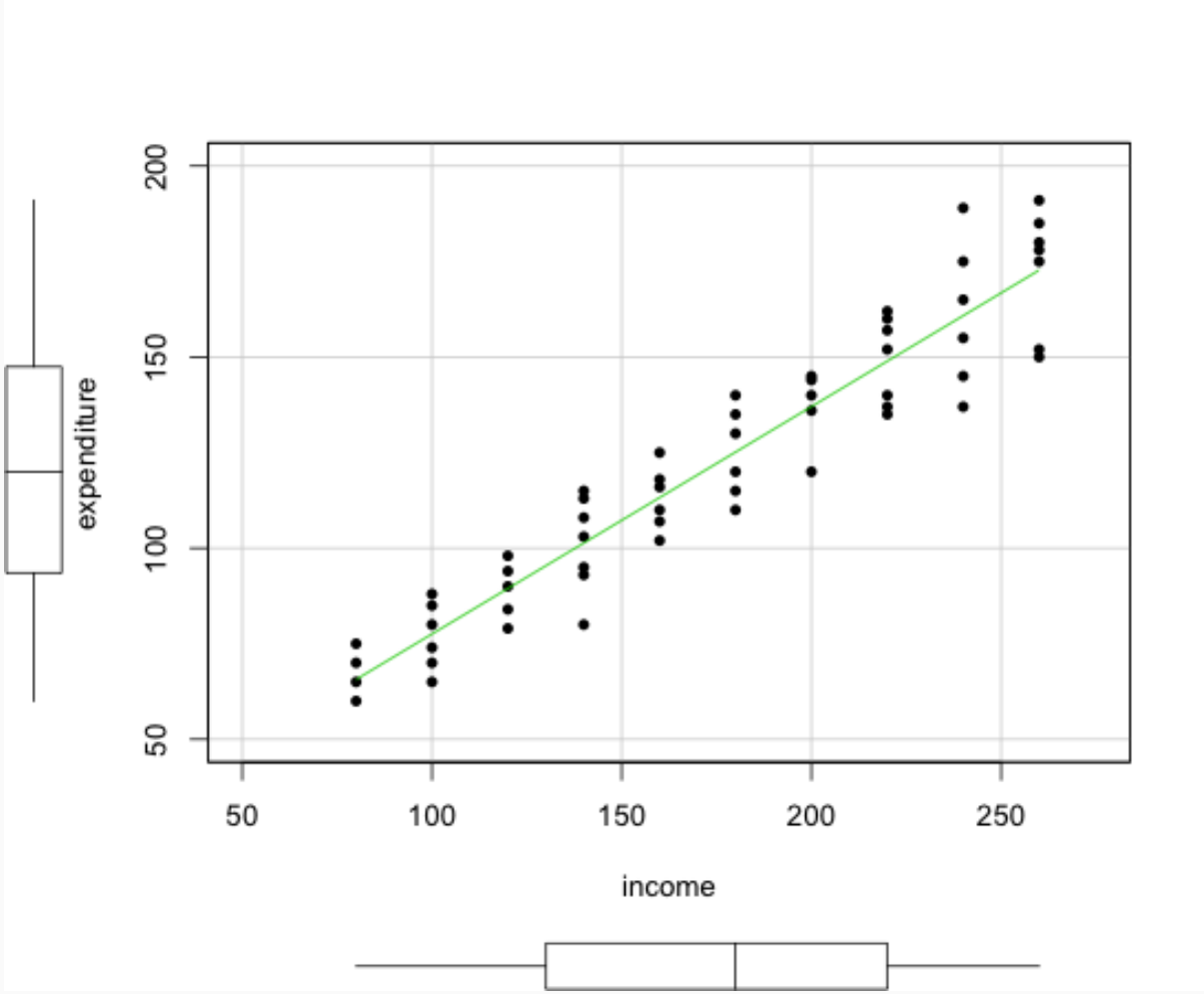
$MS_{residual}$ is the Residual standard error and is typically used as a measure of model fit (it tells us how far off the true y we would be if we used our model to predict y)

Population versus Sample Regression Function

Population Regression Function: $y = \alpha + \beta(x) + \varepsilon$

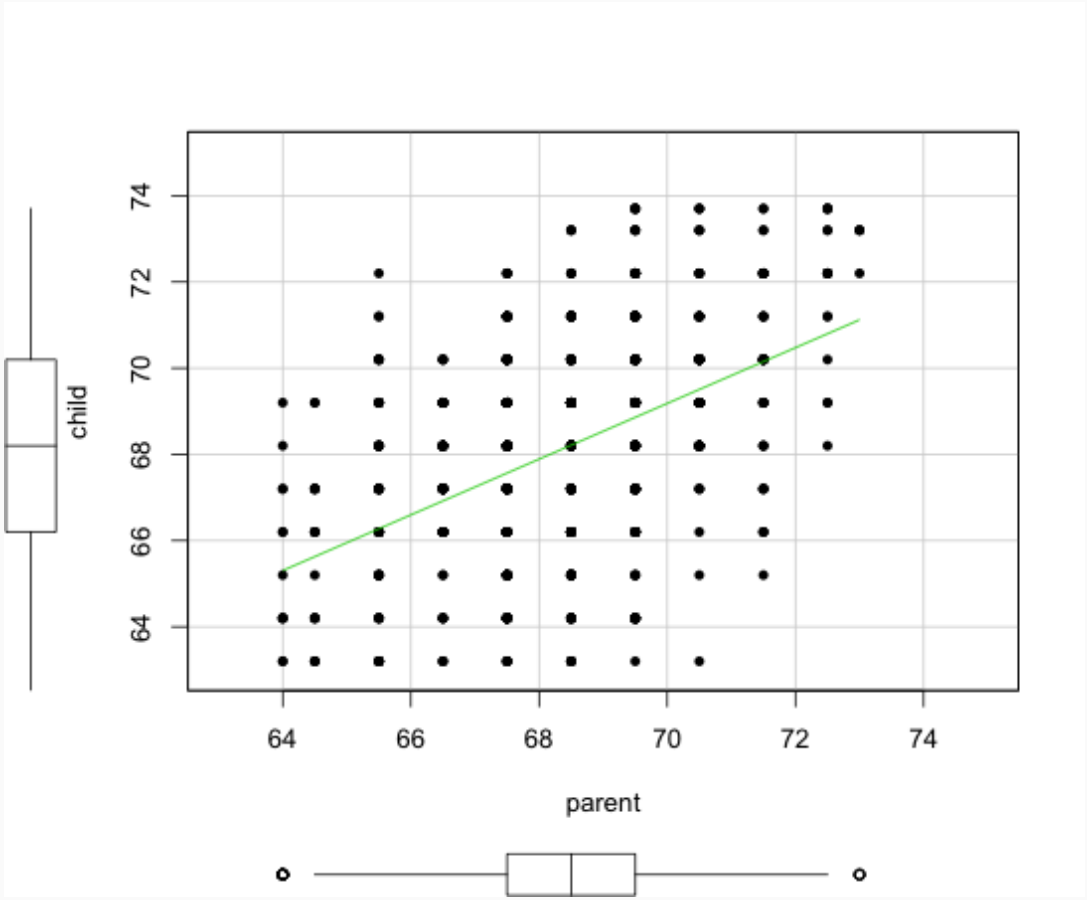
Sample Regression Function: $y = a + b(x) + e$

See the plot below: Range of y values for each *fixed* value of x_i

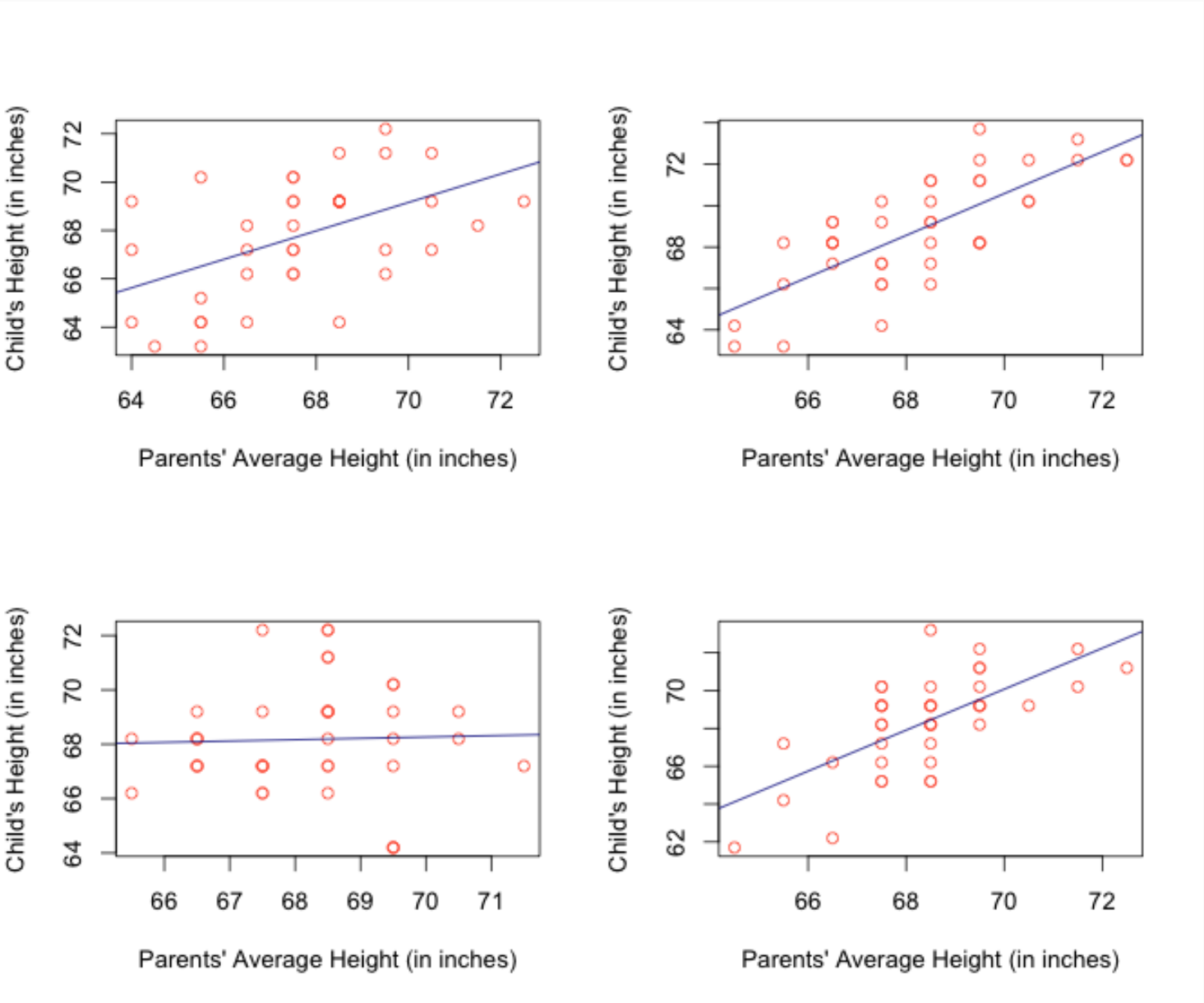


Galton's Data

These are data from a famous 1885 study of Francis Galton exploring the relationship between the heights of children and the height of their parents. The variables are the height of the adult child and the midparent height, defined as the average of the height of the father and 1.08 times the height of the mother. The units are inches. The number of cases is 928, representing 928 children and their 205 parents.



Four Sample Regression Functions



The Estimates ...

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.94153   2.81088  8.517  <2e-16 ***
Galton$parent 0.64629   0.04114 15.711  <2e-16 ***
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1        1

Sample 1
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.9453   11.1313  2.511 0.016430 *
sample1$parent 0.5888    0.1644  3.582 0.000955 ***

Sample 2
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.01339   9.62646  0.001   0.999
sample2$parent 1.00804   0.14094  7.152 1.53e-08 ***

Sample 3
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.82437  16.01798  4.047 0.000246 ***
sample3$parent 0.04915   0.23491  0.209 0.835393

Sample 4
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.7532   13.3912 -0.430    0.67
sample4$parent 1.0832    0.1958  5.532 2.49e-06 ***
```

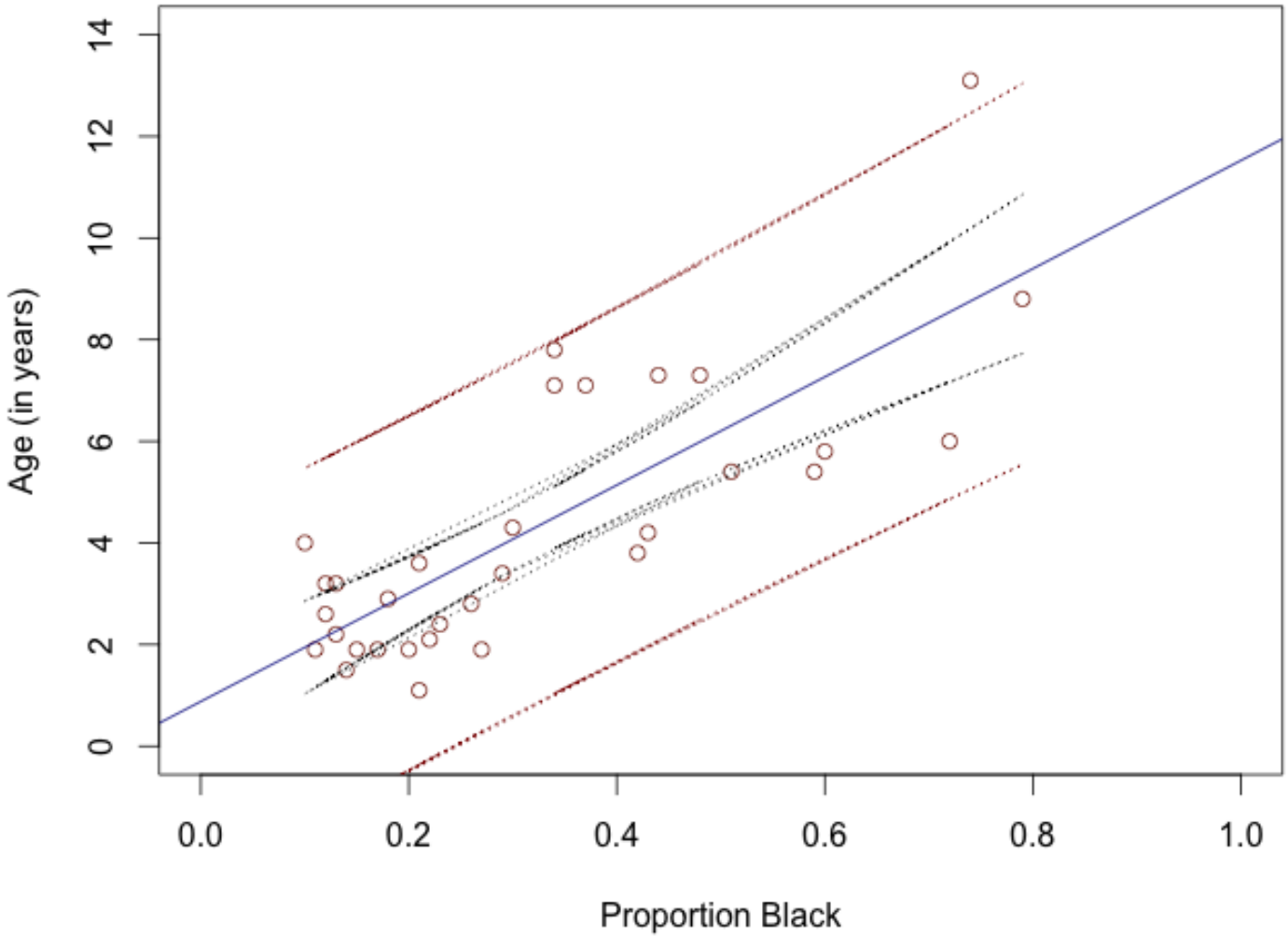
Confidence & Prediction Intervals

Estimation and Prediction

- $\hat{y}_i$ is a point estimate for y_i
- But point estimates of predicted values tell us nothing about their precision
- **Confidence intervals** and **Prediction intervals**, however, do
- Confidence interval: Interval estimate of mean value of y for specific value of x
- Prediction interval: Interval estimate of predicted value of y for specific value of x
- $x_p =$ specific value of x; $y_p =$ specific value of y for $x = x_p$
- $E(y_p) =$ expected value of y given $x = x_p$ is $\hat{y}_p = \hat{a} + \hat{b}(x_p)$
- $\text{var}(\hat{y}_p) = s^2_{\hat{y}_p} = s^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$; $\text{s}(\hat{y}_p) = s_{\hat{y}_p} = s \sqrt{\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]}$
- Confidence Interval: $\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$

- In Confidence Intervals, we could be wrong on two counts – $b_0; b_1$
- But, if we want to predict y for some x value not in the sample, we could be wrong on three counts – $b_0; b_1; e$
- So we adjust Variance as $s^2_{ind} = s^2 + s^2_{\hat{y}_p}$
- $s^2_{ind} = s^2 + s^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] = s^2 \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$
- $s_{ind} = s \sqrt{\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]}$
- Prediction Interval: $\hat{y}_p \pm t_{\alpha/2} s_{ind}$

Confidence and Prediction Intervals for LionNoses



Predicting Out of Sample Values

```
> summary(LionNoses)
      age        proportion.black
 Min.   : 1.100  Min.   :0.1000
 1st Qu.: 2.175  1st Qu.:0.1650
 Median : 3.500  Median :0.2650
 Mean   : 4.309  Mean   :0.3222
 3rd Qu.: 5.850  3rd Qu.:0.4325
 Max.   :13.100  Max.   :0.7900
> newdata <- data.frame(proportion.black=c(0.01, 0.05, 0.85, 0.90, 0.95, 0.99))
> predict(lm1, newdata, interval="predict")
        fit       lwr      upr
1  0.9854774 -2.606758 4.577713
2  1.4113622 -2.149819 4.972543
3  9.9290577 6.104725 13.753390
4 10.4614137 6.568998 14.353829
5 10.9937697 7.028446 14.959094
6 11.4196545 7.392705 15.446604
```

Multiple Linear Regression

Multiple Linear Regression

Multiple Regression Analysis refers to models with more than one independent variable as in, for example, $y = a + b_1(x_1) + b_2(x_2) + b_3(x_3) + e$

Basic interpretation remains the same except the slopes $b_1, b_2, b_3,$ etc. are referred to as the partial slope coefficients (because no single variable explains the slope of the regression line on its own)

Allows for all categorical variables, all continuous variables, or a mix of the two types

The data used below reflect the number of deaths in London from 1st-15th Dec 1952 due to air pollution. Two independent variables are usable – atmospheric smoke (in mg/cu.m), and SO2 (atmospheric sulphur dioxide in parts/million). The dependent variable will be the number of deaths.

```
Call:
lm(formula = deaths ~ smoke + SO2, data = SO2)
Residuals:
    Min      1Q  Median      3Q     Max
-100.717 -20.689  -3.298  15.148 114.931
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   89.51     25.08   3.569 0.003858 **
smoke       -220.32     58.14  -3.789 0.002579 **
SO2         1051.82    212.60   4.947 0.000338 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
Residual standard error: 52.96 on 12 degrees of freedom
Multiple R-squared: 0.859,   Adjusted R-squared: 0.8355
F-statistic: 36.57 on 2 and 12 DF, p-value: 7.844e-06
```

Understanding the Estimates

$y = 89.51 - 220.32(smoke) + 1051.82(SO2)$

You see the (Intercept) plus partial slope coefficients for smoke and SO2

The (Intercept) is interpreted as follows: It is the predicted number of deaths when smoke $= 0$ and SO2 $= 0$

The partial slope coefficient on smoke is -220.32 and indicates that holding all else constant (which in this case means holding SO2 constant) every unit increase (i.e., an increase of 1) in smoke decreases the number of deaths by about 220

The partial slope coefficient on SO2 is 1051.82 and indicates that holding all else constant (which in this case means holding smoke constant) every unit increase in SO2 increases the number of deaths by about 1052

The adjusted R^2 is 0.8355, indicating that about 83.55% of the variation in the number of deaths can be jointly explained by smoke and SO2

The Residual standard error is 52.96, indicating that if we try to predict the number of deaths using this regression model, on average our prediction would be off the TRUE number of deaths by about 53 $\cdots$ that is, 53 more than really occur or 53 less than really occur

Predicting the Number of Deaths

We can calculate predicted number of deaths by using the estimated regression model but in order to do so we will need to specify values for smoke and for SO2

One good way to show how the model works is by choosing specific values of the independent variables when generating predictions. Most commonly one would pick the minimum, first quartile, mean (or median), third quartile, and the maximum values of each independent variable

Below we do this:

```
> summary(SO2)
      day           deaths         smoke           SO2
 Min.   : 1.0   Min.   :112.0   Min.   :0.290   Min.   :0.090
 1st Qu.: 4.5   1st Qu.:169.5   1st Qu.:0.320   1st Qu.:0.160
 Median : 8.0   Median :236.0   Median :0.500   Median :0.230
 Mean   : 8.0   Mean   :261.5   Mean   :1.406   Mean   :0.458
 3rd Qu.:11.5   3rd Qu.:284.0   3rd Qu.:1.930   3rd Qu.:0.610
 Max.   :15.0   Max.   :518.0   Max.   :4.460   Max.   :1.340

> new.data = data.frame(smoke = c(0.290, 0.320, 0.500, 1.930, 4.460), SO2 = c(0.090, 0.160, 0.230, 0.610, 1.340))
> yhat = predict(lm.SO2b, newdata=new.data, interval="conf")
> yhat
       fit      lwr      upr
1 120.2802 71.98191 168.5785
2 187.2976 150.41784 224.1774
3 221.2664 185.58679 256.9460
4 305.8928 273.95491 337.8307
5 516.2981 443.57961 589.0167
```

```
> new.data = data.frame(smoke = c(0.290, 0.320, 0.500, 1.930, 4.460), SO2 = c(0.090, 0.160, 0.230, 0.610, 1.340))
> yhat = predict(lm.SO2b, newdata=new.data, interval="conf")
> yhat
      fit      lwr     upr
1 120.2802 71.98191 168.5785
2 187.2976 150.41784 224.1774
3 221.2664 185.58679 256.9460
4 305.8928 273.95491 337.8307
5 516.2981 443.57961 589.0167
```

Note that row 5 is for smoke = 4.460 and SO2 = 1.340 – i.e., both at their maximum

row 3 is when both are at their in-sample median values

But you can tweak these combinations as you want to so long as you do not step outside the in-sample values of your independent variables

For example, let us see how SO2 impacts deaths when it is at its maximum and smoke is at its minimum. We can then reverse this specification

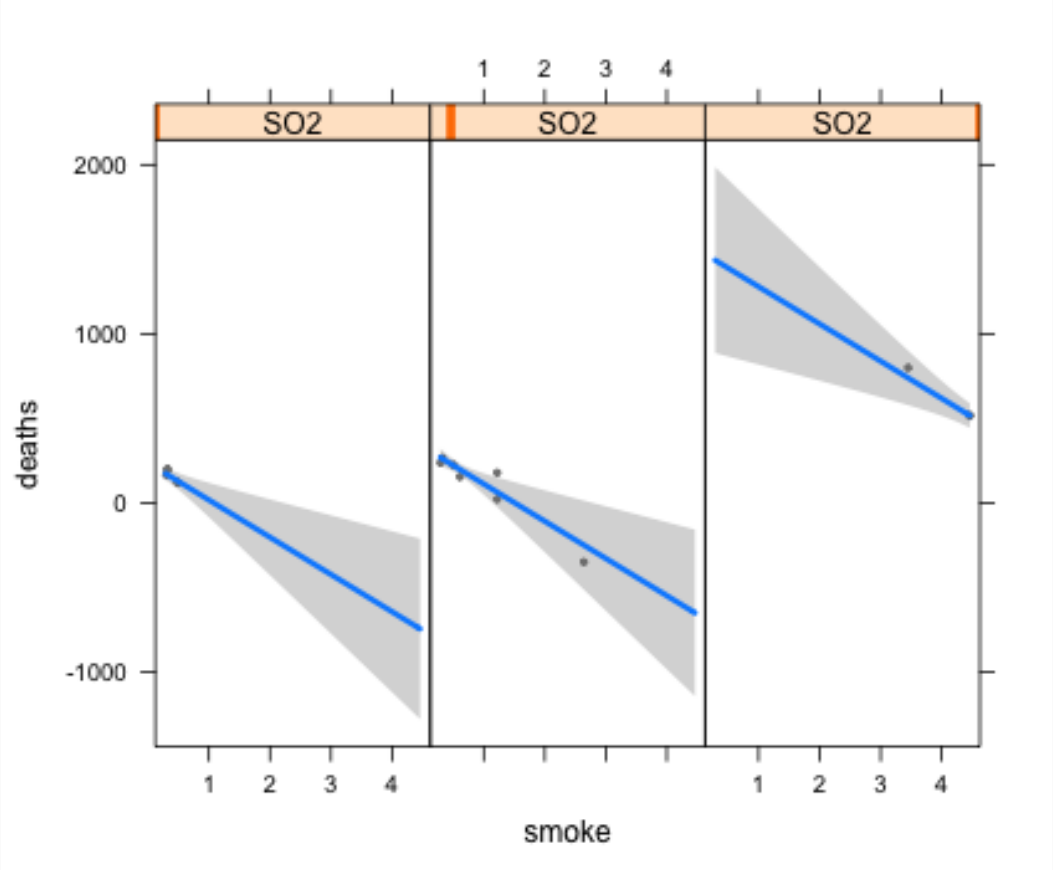
```
> new.data = data.frame(smoke = 0.290, SO2 = 1.340)
> yhat = predict(lm.SO2b, newdata=new.data, interval="conf")
> yhat
      fit     lwr     upr
1 1435.051 885.6235 1984.478
> new.data = data.frame(smoke = 4.460, SO2 = 0.090)
> yhat = predict(lm.SO2b, newdata=new.data, interval="conf")
> yhat
       fit      lwr     upr
1 -798.4724 -1355.147 -241.798
```

Note: The latter predictions make no sense since you cannot have negative deaths!!

Plotting the Regression Line

An easy way to demonstrate the results is by plotting the predicted values of the dependent variable according to values of the independent variables via visreg

```
visreg(lm.SO2b, "smoke", by="SO2")
```



SO2 is being held constant at the 10^{th}, 50^{th}, and 90^{th} percentiles

Categorical Independent Variables

Treatments as Categorical Independent Variables

- Thus far we have looked at x as a continuous independent variable but what if the independent variable is categorical, say a "treatment" with two or more levels?
- For example, recall Chapter 15's LodgepolePines data where the response variable was the mean cone size recorded in 16 sites that fell into three mutually exclusive categories – Island without squirrels; Island with squirrels; and Mainland with squirrels
- We can easily fit a regression model to these data as well, but with some modifications in terms of how the treatment indicator enters the model
  We may want to specify the model as
  $y = a + b_1(Island.present) + b_2(Mainland.present) + b_3(Island.absent)$
  ... but this would be a problem because in the regression setting the intercept represents the expected value of y when $x = 0$. Here, that would imply a site that is not an Island with squirrels present, nor is it the Mainland with squirrels present, nor an Island with squirrels absent ... i.e., a site that does not exist in the study!!
  So we *always* exclude one treatment condition (typically the one with largest n_j) and use it as the reference category
  The model then might be $y = a + b_1(Island.present) + b_2(Mainland.present)$

$y = a + b_1(Island.present) + b_2(Mainland.present)$

If the site is Island.absent, then the equation becomes

$y = a + b_1(Island.present = 0) + b_2(Mainland.present = 0) = a$

If the site is Island.present, then the equation becomes

$y = a + b_1(Island.present = 1) + b_2(Mainland.present = 0) = a + b_1$

If the site is Mainland.present, then the equation becomes

$y = a + b_1(Island.present = 0) + b_2(Mainland.present = 1) = a + b_2$

... so expected mean conemass is a for Island.absent, $a + b_1$ for Island.present, and $a + b_2$ for Mainland.present

```
Call:
lm(formula = conemass ~ habitat, data = LodgepolePines)
Residuals:
   Min     1Q Median     3Q    Max
-0.780 -0.405 -0.040 0.505 0.720
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             8.9000     0.2212  40.238 4.97e-15 ***
habitatisland present  -2.8200     0.3281  -8.596 1.01e-06 ***
habitatmainland present -2.7800    0.3281  -8.474 1.18e-06 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
Residual standard error: 0.5418 on 13 degrees of freedom
Multiple R-squared: 0.8851,  Adjusted R-squared: 0.8675
F-statistic: 50.09 on 2 and 13 DF, p-value: 7.787e-07
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             8.9000     0.2212  40.238 4.97e-15 ***
habitatisland present  -2.8200     0.3281  -8.596 1.01e-06 ***
habitatmainland present -2.7800    0.3281  -8.474 1.18e-06 ***
```

mean cone mass is predicted to be 8.90 for Island.absent, $8.90 - 2.82 = 6.08$ for Island.present, and $8.90 - 2.78 = 6.12$ for Mainland.present

Now look at the mean cone mass for each habitat type (see below)

```
> with(pines, tapply(conemass, habitat, mean))
   island.absent  island.present mainland.present
            8.90            6.08            6.12
```

Note: 8.90 - 6.08 = -2.82 ... average cone mass is lower for island.present than that for island.absent by 2.82
Note: 8.90 - 6.12 = -2.78 ... average cone mass is lower for mainland.present than that for island.absent by 2.78
So the regression estimates, the partial slope coefficients here, give you the difference between each group represented by the categorical variables
The intercept is the mean for the reference group (here Island.absent)

In general, with m categories of the variable you will see estimates for $m - 1$ categories

Two Categorical Independent Variables

Recall the flies data we used for two-factor ANOVA

The dependent variable was lifespan (in days), and the two independent variables were (1) fertility (fertile vs. sterile), and (2) treatment (low-cost vs. high-cost)

The linear model to be estimated is $lifespan = a + b_1(fertility) + b_2(treatment) + e$

```
> lm.f = lm(lifespanDays ~ treatment + fertility, data=flies)
> summary(lm.f)

Call:
lm(formula = lifespanDays ~ treatment + fertility, data = flies)
Residuals:
    Min      1Q  Median      3Q     Max
-30.3061 -4.3290  0.8528  4.8528 25.8757
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      23.1472    0.4926 46.988 < 2e-16 ***
treatmentlow-cost 6.9771    0.5684 12.276 < 2e-16 ***
fertilitysterile  2.1819    0.5684  3.839 0.000133 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
Residual standard error: 8.246 on 839 degrees of freedom
Multiple R-squared: 0.1649,  Adjusted R-squared: 0.1629
F-statistic: 82.83 on 2 and 839 DF, p-value: < 2.2e-16
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      23.1472    0.4926 46.988 < 2e-16 ***
treatmentlow-cost 6.9771    0.5684 12.276 < 2e-16 ***
fertilitysterile  2.1819    0.5684  3.839 0.000133 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
Residual standard error: 8.246 on 839 degrees of freedom
Multiple R-squared: 0.1649,  Adjusted R-squared: 0.1629
F-statistic: 82.83 on 2 and 839 DF, p-value: < 2.2e-16
```

R will use the first label of the factor as the reference category (high-cost comes before low-cost and fertile comes before sterile) If we were interested in the high-cost + fertile group, both of these partial slopes (treatmentlow-cost and sterile) would be dropped, which would make the (Intercept) the estimated lifespan for the high-cost + fertile group (the reference group now)

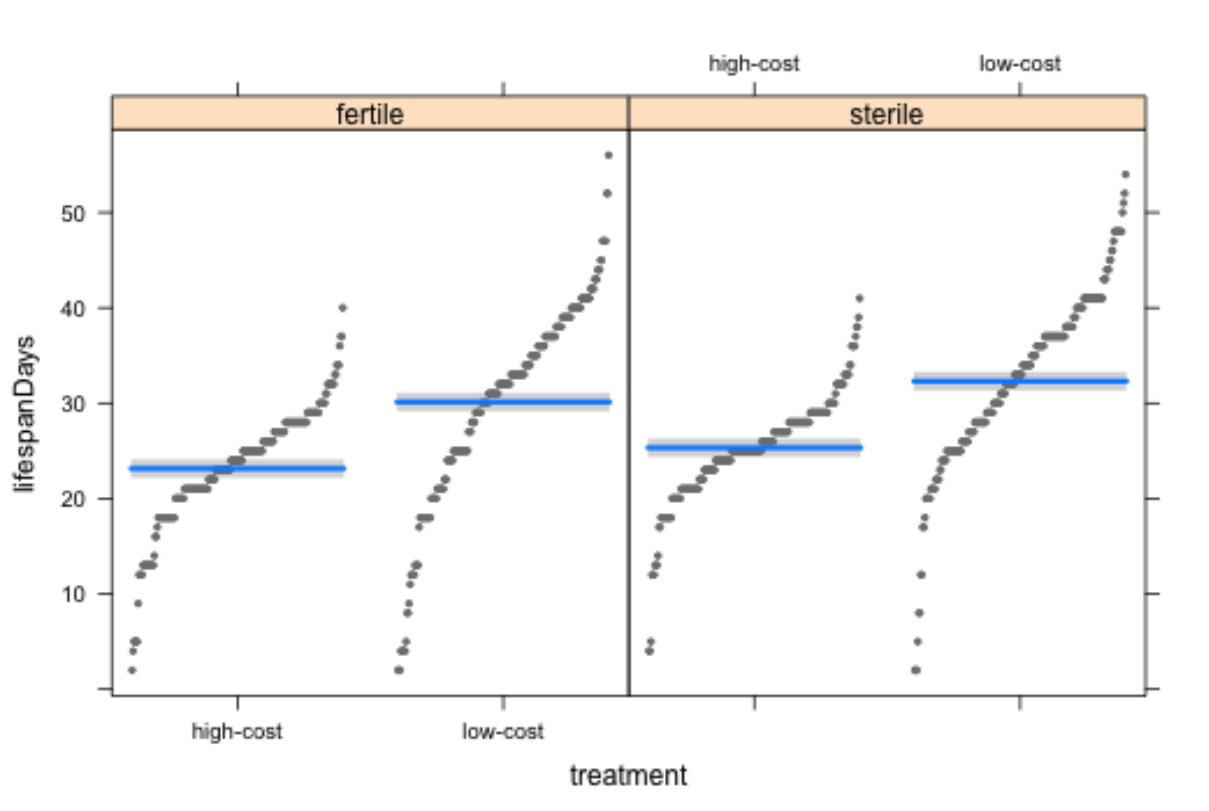
If we wanted the estimated lifespan for the low-cost + fertile group then the partial slope on fertilitysterile would be dropped

If we wanted the the estimated lifespan for the high-cost + sterile group then the partial slope on treatmentlow-cost would be dropped

Predicting the mean lifespan for particular groups ...

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      23.1472     0.4926  46.988  < 2e-16 ***
treatmentlow-cost 6.9771     0.5684  12.276  < 2e-16 ***
fertilitysterile  2.1819     0.5684   3.839 0.000133 ***
```

for high-cost + fertile group: $= \hat{a} = 23.1472$
for high-cost + sterile group: $= \hat{a} + \hat{b}_2 = 23.1472 + 2.1819 = 25.3291$
for low-cost + fertile group: $= \hat{a} + \hat{b}_1 = 23.1472 + 6.9771 = 30.1243$
for low-cost + sterile group: $= \hat{a} + \hat{b}_1 + \hat{b}_2 = 23.1472 + 6.9771 + 2.1819 = 32.3062$



Mixing Variable Types: Mole Rats

Let us now fit a model with two independent variables – one continuous and one categorical

Mole rats are the only known mammals with distinct social castes. A single queen and a small number of males are the only reproducing individuals in a colony. Remaining individuals, called workers, gather food, defend the colony, care for the young, and maintain burrows. Recently, it was discovered that there might be two worker castes in the Damaraland mole rat. “Frequent workers” do almost all of the work in the colony, whereas ”infrequent workers” do little work except on rare occasions after rains, when they extend the burrow system. To assess the physiological differences between the two types of workers, researchers compared daily energy expenditures of wild mole rats during a dry season. Energy expenditure appeared to vary with body mass in both groups but infrequent workers were heavier than frequent workers. How different is mean daily energy expenditure between the two groups when adjusted for differences in body mass?

We will fit the following model:

$$log(energy) = a + b_1(log(mass)) + b_2(caste) + e$$

```
> lm.r = lm(lnEnergy ~ lnMass + caste, data=rats)
> summary(lm.r)
Call:
lm(formula = lnEnergy ~ lnMass + caste, data = rats)
Residuals:
    Min      1Q  Median      3Q     Max
-0.73388 -0.19371 0.01317 0.17578 0.47673
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.09687  0.94230 -0.103  0.9188
lnMass      0.89282   0.19303  4.625 5.89e-05 ***
casteworker 0.39334   0.14611  2.692  0.0112 *
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
Residual standard error: 0.2966 on 32 degrees of freedom
Multiple R-squared: 0.409,   Adjusted R-squared: 0.3721
F-statistic: 11.07 on 2 and 32 DF, p-value: 0.0002213
```

Note: both lnMass and caste are statistically significant
The (Intercept) is predicted lnEnergy when caste=lazy and lnMass = 0 ... makes no sense for lnMass but that is typically what happens with intercepts
Predicted lnEnergy for a worker is $= -0.09687 + 0.82928(lnMass) + 0.39334$ and we'll have to set a value of lnMass to calculate the final result

Typically one would calculate the predicted value of the dependent variable by picking substantively interesting values of the independent variable. In our case, say we use the five-number summary – Min., Q_1, Median, Q_3, and Max.

```
> summary(rats$lnEnergy)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  3.555   3.902  4.190  4.193  4.489  5.043
```

We'll create a new data-frame that contains these values and generate predictions, one set for workers and the other for lazy:

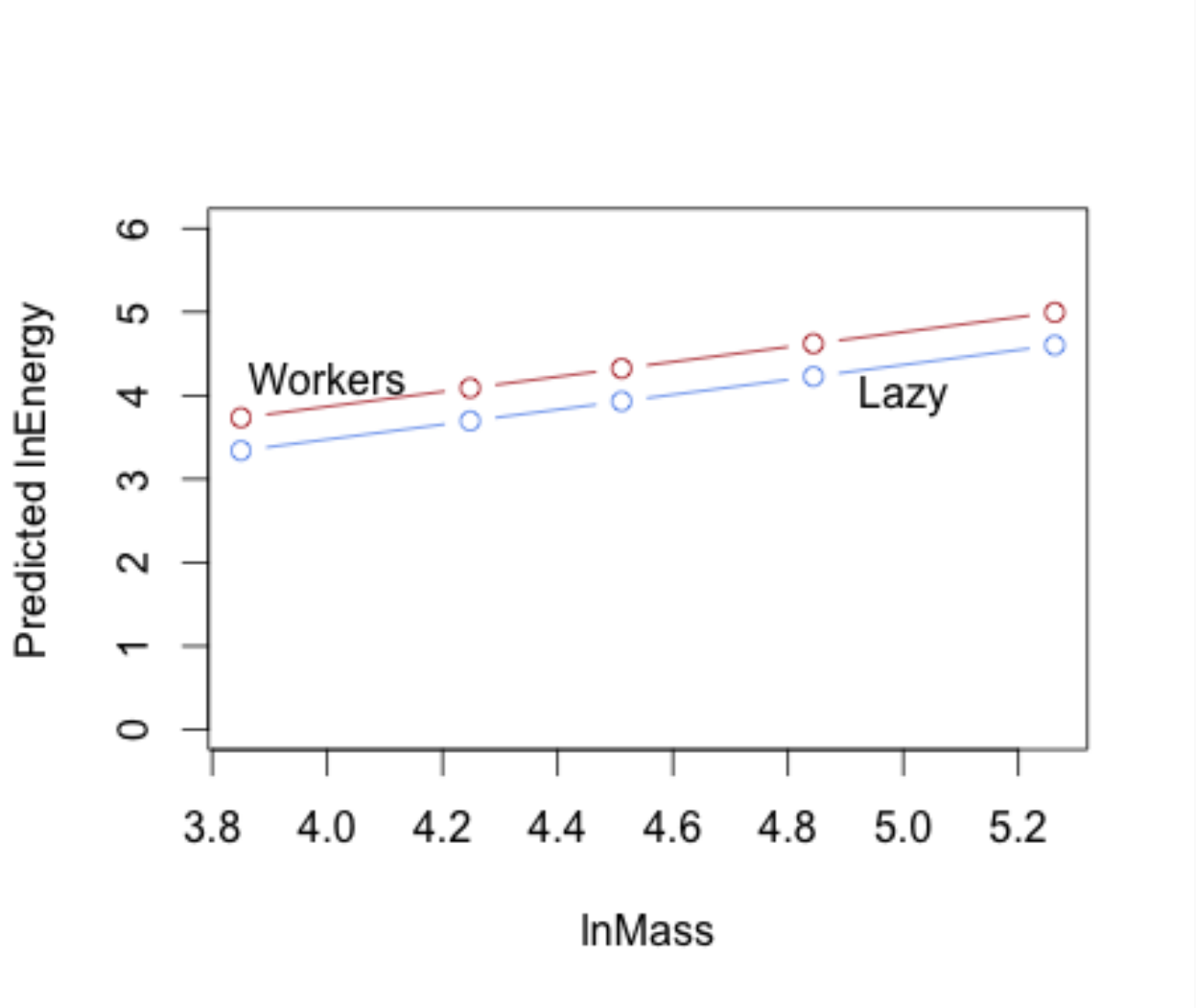
```
> new.data.a = data.frame(lnMass=c(3.850, 4.248, 4.511, 4.844, 5.263 ), caste="worker")
> new.data.b = data.frame(lnMass=c(3.850, 4.248, 4.511, 4.844, 5.263 ), caste="lazy")
> predicted.lnEnergy.w = predict(lm.r, newdata=new.data.a)
> predicted.lnEnergy.l = predict(lm.r, newdata=new.data.b)
> predicted.lnEnergy.w
      1       2       3       4       5
3.733812 4.089153 4.323963 4.621270 4.995360
> predicted.lnEnergy.l
      1       2       3       4       5
3.340470 3.695810 3.930621 4.227928 4.602018
```

Notice the difference between the predictions at each value of lnMass we specified:

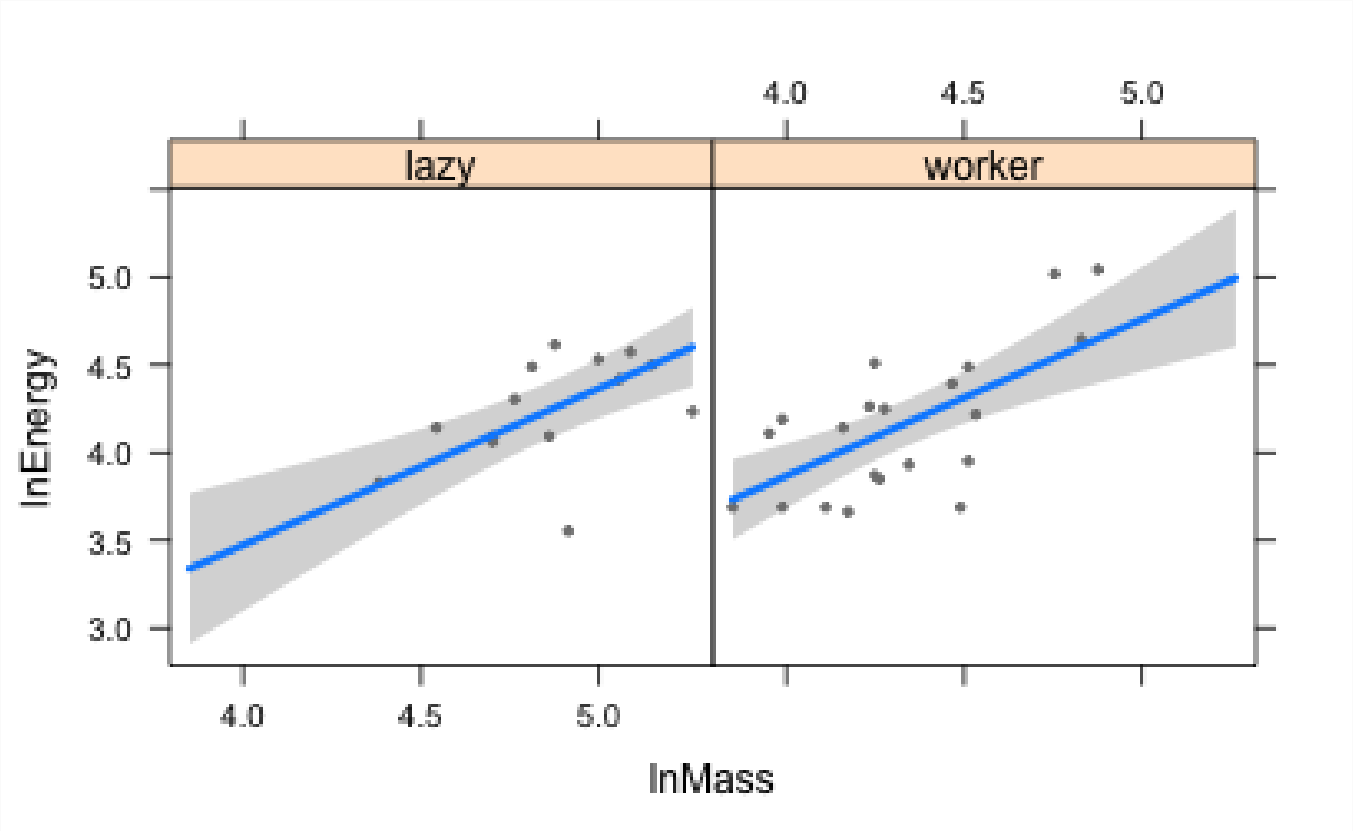
```
> predicted.lnEnergy.w - predicted.lnEnergy.l
       1        2        3        4        5
0.3933424 0.3933424 0.3933424 0.3933424 0.3933424
```

... it is the partial slope coefficient for caste

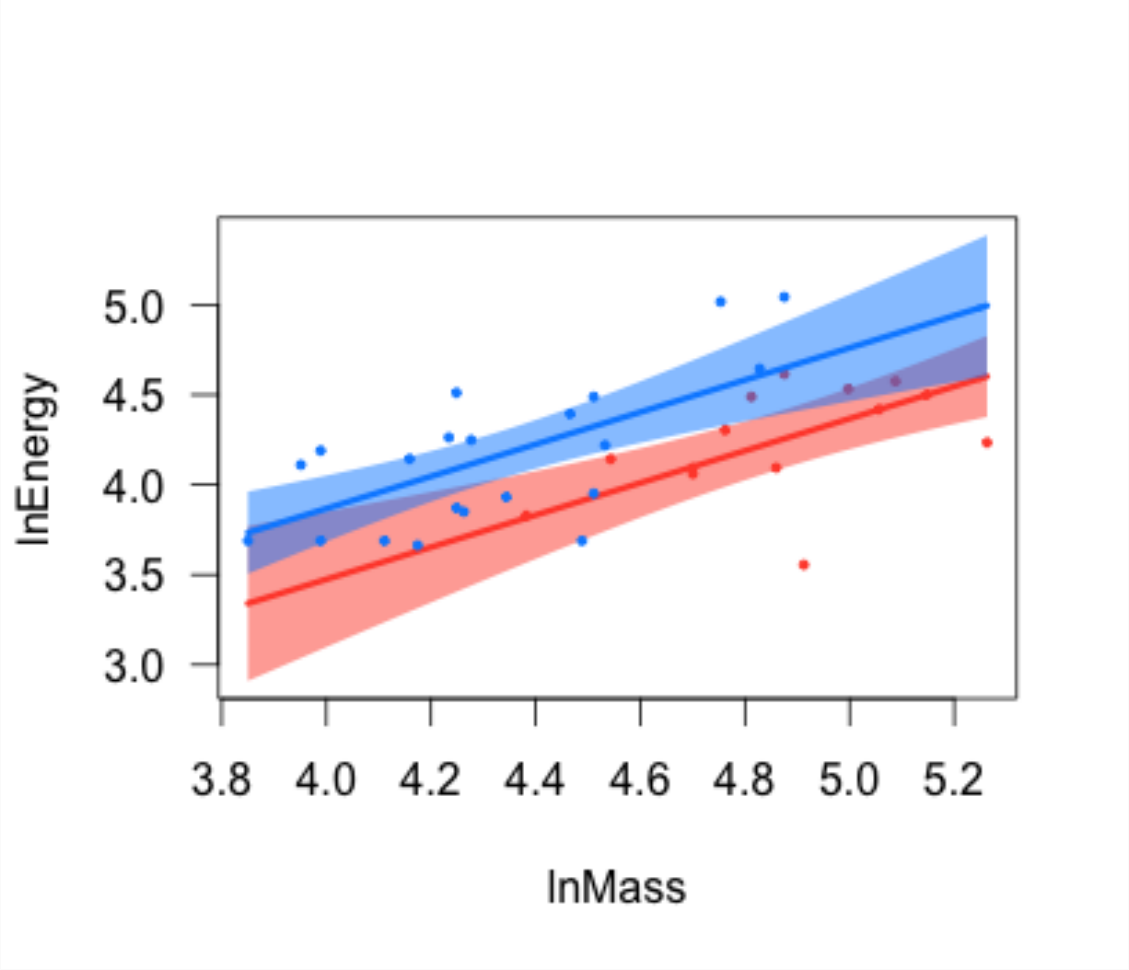
If we plot the two sets of predictions you'll see this constant difference showing up as an intercept-shift (upwards for workers)



Plotting with visreg and using in-sample values of lnMass



Plotting with visreg and using in-sample values of lnMass but overlaying the two panels



The F test

Say we have a model such as the following: $y = a + b_1(x_1) + b_2(x_2)$

How can we test that both b_1 and b_2 are not simultaneously zero?

... via the F test where $F = \frac{\frac{SSR}{k-1}}{\frac{SSE}{n-k}}$ where $k =$ number of independent variables, SSR is Sum of Squares due to the Regression, and SSE is Sum of Squares due to the Errors

Here $H_0 : b_1 = b_2 = \ldots = b_k = 0$ and H_A : Not all b_i are simultaneously equal to 0

R automatically gives you this F statistic and the associated p-value

If the p-value ≤ 0.05 we can reject H_0

```
> lm.SO2b = lm(deaths ~ smoke + SO2, data=SO2)
> summary(lm.SO2b)
Call:
lm(formula = deaths ~ smoke + SO2, data = SO2)
Residuals:
    Min      1Q  Median      3Q     Max
-100.717 -20.689  -3.298  15.148 114.931
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   89.51     25.08   3.569 0.003858 **
smoke       -220.32     58.14  -3.789 0.002579 **
SO2         1051.82    212.60   4.947 0.000338 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
Residual standard error: 52.96 on 12 degrees of freedom
Multiple R-squared: 0.859,   Adjusted R-squared: 0.8355
F-statistic: 36.57 on 2 and 12 DF, p-value: 7.844e-06
```

Note: F-statistic: 36.57 on 2 and 12 DF, p-value: 7.844e-06

The Marginal Contribution of an Independent Variable

Typically you will have two or more independent variables you want to test in a regression

This often raises the question of whether adding one or more variables really adds much in terms of improving the fit of the model to the data

The F test can be used here as well to decide if the "new" model is an improvement over the "old" model

$$F = \frac{\frac{SSR_{new} - SSR_{old}}{\text{number of new independent variables}}}{\frac{SSE_{new}}{n - k_{new} - 1}}$$

This is the same as running $F = \frac{R^2_{new} - R^2_{old} / \text{number of new independent variables}}{\left(1 - R^2_{new}\right) / n - k_{new} - 1}$

Let us see this test in action with respect to the SO2 data

```
> lm.SO2a = lm(deaths ~ smoke, data=SO2)
> summary(lm.SO2a)
Call:
lm(formula = deaths ~ smoke, data = SO2)
Residuals:
    Min      1Q  Median      3Q     Max
-144.15  -73.33   24.39   54.55  180.39
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  171.82      31.43   5.466 0.000108 ***
smoke         63.76      15.31   4.164 0.001112 **
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
Residual standard error: 88.71 on 13 degrees of freedom
Multiple R-squared: 0.5715,  Adjusted R-squared: 0.5386
F-statistic: 17.34 on 1 and 13 DF, p-value: 0.001112
> lm.SO2b = lm(deaths ~ smoke + SO2, data=SO2)
> summary(lm.SO2b)
Call:
lm(formula = deaths ~ smoke + SO2, data = SO2)
Residuals:
     Min       1Q   Median       3Q      Max
-100.717  -20.689   -3.298   15.148  114.931
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   89.51      25.08   3.569 0.003858 **
smoke       -220.32      58.14  -3.789 0.002579 **
SO2         1051.82     212.60   4.947 0.000338 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
Residual standard error: 52.96 on 12 degrees of freedom
Multiple R-squared: 0.859,   Adjusted R-squared: 0.8355
F-statistic: 36.57 on 2 and 12 DF, p-value: 7.844e-06
```

$R^2_{lm.SO2b} = 0.8590$ and $R^2_{lm.SO2a} = 0.5715$

Number of new independent variables $= 1$ and $k_{new} = 2$

Therefore, $F = \frac{(0.8590 - 0.5715)/1}{(1 - 0.8590)/(15 - 3)} \approx 24.46809$, which turns out to have a very low p-value

Conclusion? SO2 should be added to the model

In R we can compare such nested models via the anova() command ...

```
> anova(lm.SO2a, lm.SO2b)
Analysis of Variance Table
Model 1: deaths ~ smoke
Model 2: deaths ~ smoke + SO2
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     13 102302
2     12  33654  1     68648 24.478 0.0003378 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
```

Before we move on, see how the effect of smoke differs in the two models

Some Cautions

- Never extrapolate beyond the in-sample values of the independent variables
- Always focus on the adjusted R^2 because it penalizes the R^2 for the number of independent variables being used. As you increase the number of independent variables the R^2 will always increase even if the new variables are not statistically significant
- Never look to maximize R^2 because even the worst models can yield very high R^2 values *even when no independent variable is statistically significant*
- Never ignore the Residual standard error because that is a good indicator of average prediction error one could make if one used the model
- Always test your regression model against new data ... how well it performs will be determined by how well it actually predicts the actual dependent variable values in the new sample
- Test for interactions (if theory or suspicions suggest as much)
- Remember Occam's Razor: "when you have two competing theories that make exactly the same predictions, the simpler one is better"

Assumptions of Linear Regression

Core Assumptions of Simple Linear Regression

1. The data you have map well to the research question you are wrestling with
2. The regression is linear in parameters ... the slope b appears with a power of 1. Thus $y = a + b\left(x^2\right)$ is still a linear regression model, as is $y = a + b\left(\frac{1}{x}\right)$, etc.
3. Each value of x is associated with a population of y values, and the mean of these y values falls on the population regression line (i.e., $E(e_i|x_i) = 0$)
4. For each value of x the population of y values is normally distributed
5. For each x value the sampled y values are a random draw from the corresponding population of y values
6. The variance of the y values is constant across all x values ... This is known as heteroscedasticity
7. You have no outliers that significantly influence the regression line
8. No two or more variables are highly correlated ... This is known as multicollinearity

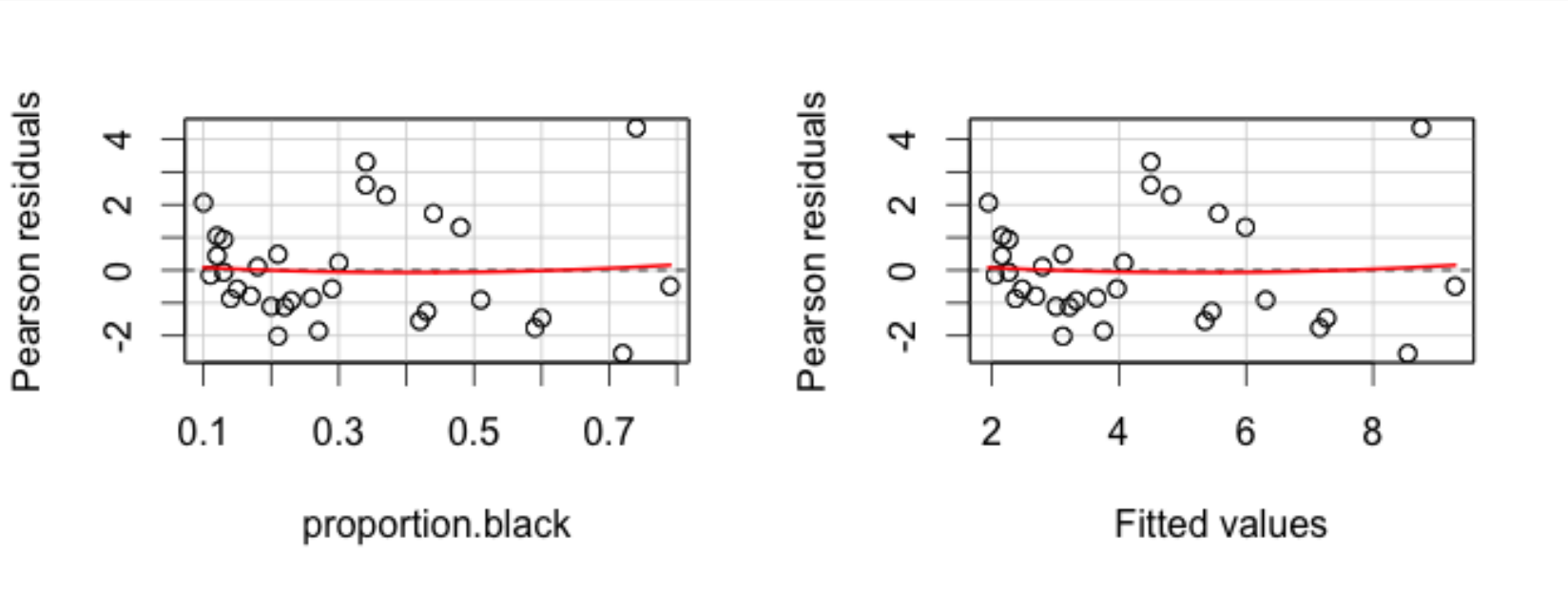
The car library has several useful diagnostics to identify violations of these (and more) regression assumptions

Testing Assumptions

- $e_i = y_i - \hat{y}_i$ is the residual, aka the prediction error for each observation i
- ordinary residuals: $e_i = y_i - \hat{y}_i$, where $i = 1, 2, 3, \ldots, n$
- Residuals are the basis of most diagnostic methods because if the regression model is *correct* then ordinary residuals should be strictly random with mean and variance given by: $E(e_i) = 0$; $Var(e_i) = \sigma^2(1 - h_{ii})$
  - h_{ii} is the *leverage* or *hat-value*, and indicates how much of an impact a particular observation has on the regression line ... a large h_{ii} is a sign of an unusual x_i value; observations close to the center of the regression space will have small h_{ii}
  - Think of h_{ii} as follows. Assume for a given value x_i you change y_i by a little bit and then re-estimate the regression. This change may not really modify the original regression, in which case the original $\hat{y}_i$ and the new $\hat{y}_i$ will be the same. On the other hand, if the new $\hat{y}_i$ is now different from its original value, well then this particular y_i is shaping the regression line a good bit. Thus $h_{ii} = \frac{\delta \hat{y}_i}{\delta y_i}$
- If plots of residuals against $\hat{y}$ and the x variable show some non-randomness, this is a sign that one or more assumptions is being violated
- In the residualPlots() that follow a flat red line $=$ all okay; any marked curvilinear pattern suggests something is wrong

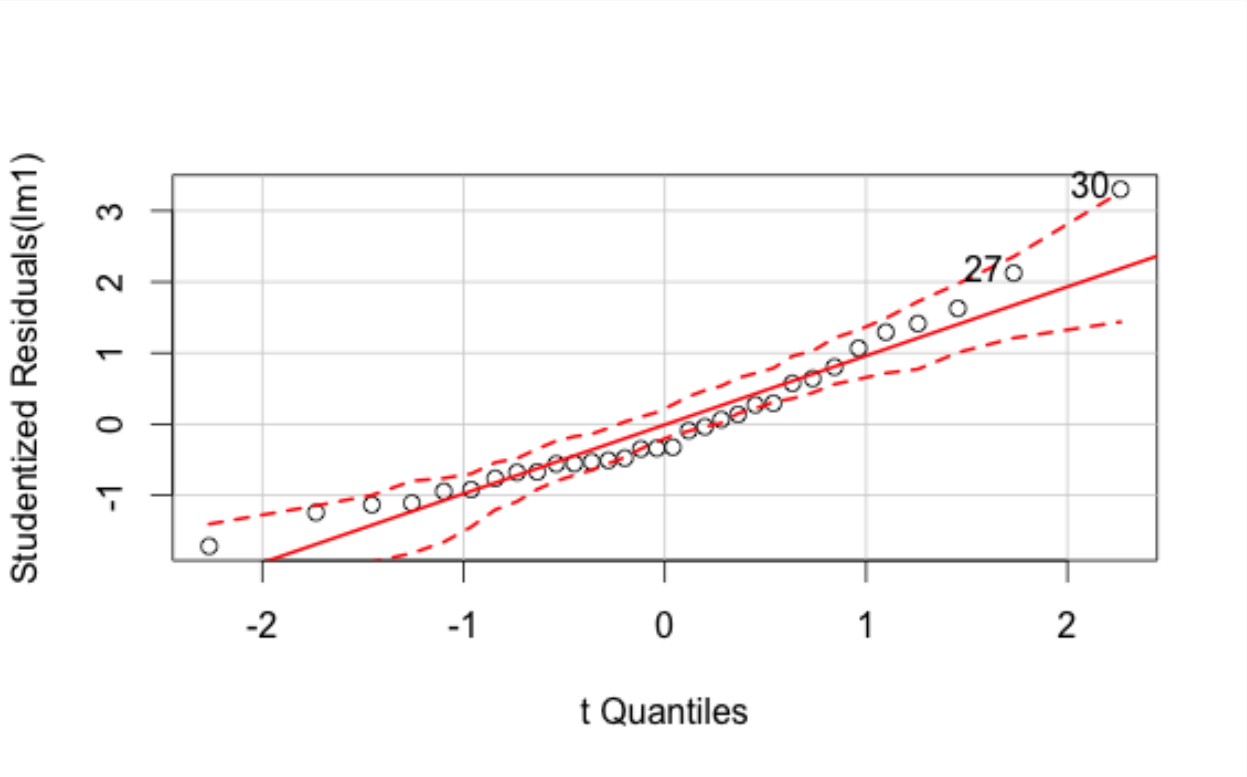
LionNoses Revisited

```
Call:
lm(formula = age ~ proportion.black, data = LionNoses)
Residuals:
    Min      1Q  Median      3Q     Max
-2.5449 -1.1117 -0.5285 0.9635 4.3421
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.8790    0.5688   1.545    0.133
proportion.black 10.6471    1.5095   7.053 7.68e-08 ***
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1        1
Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared: 0.6238,  Adjusted R-squared: 0.6113
F-statistic: 49.75 on 1 and 30 DF, p-value: 7.677e-08
```



Outliers and Studentized Residuals

- If an observation is having an unusually large effect on the regression model we should be able to model it via $y_i = a + b(x_i) + \gamma(d_i)$, where $d_i = 1$ for observation i and $d_i = 0$ for all other observations
- Studentized residuals essentially perform such a test
  Studentized Residual: $e_{\tau i} = \dfrac{e_i}{\hat{\sigma}_{-i}\sqrt{1-h_{ii}}}$; note that $\hat{\sigma}_{-i}$ is the estimated variance of the residuals when observation i has been excluded and the regression model fit again. This is akin to the following:
  1. Run the full regression model with all observations and save the residuals
  2. Now exclude each observation – one at a time – and refit the model, saving the residuals at each run
  3. Calculate the ratio of the full model residual of observation i to the standard deviation of the residuals from the regression excluding observation i
  4. This ratio follows the t distribution with $n-k-2$ degrees of freedom
  5. Multiple simultaneous tests so we use Bonferroni adjustment of the P-value



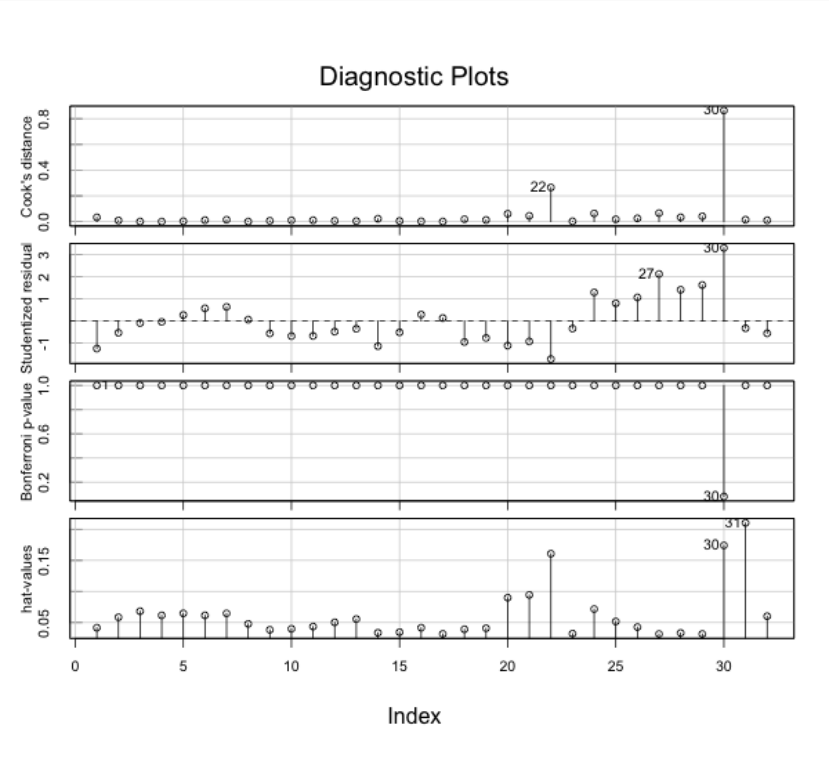
Note that two observations with the largest absolute Studentized residuals are flagged – observations 27 and 30, but both are within the 95% confidence interval so we may have no problem here with outliers

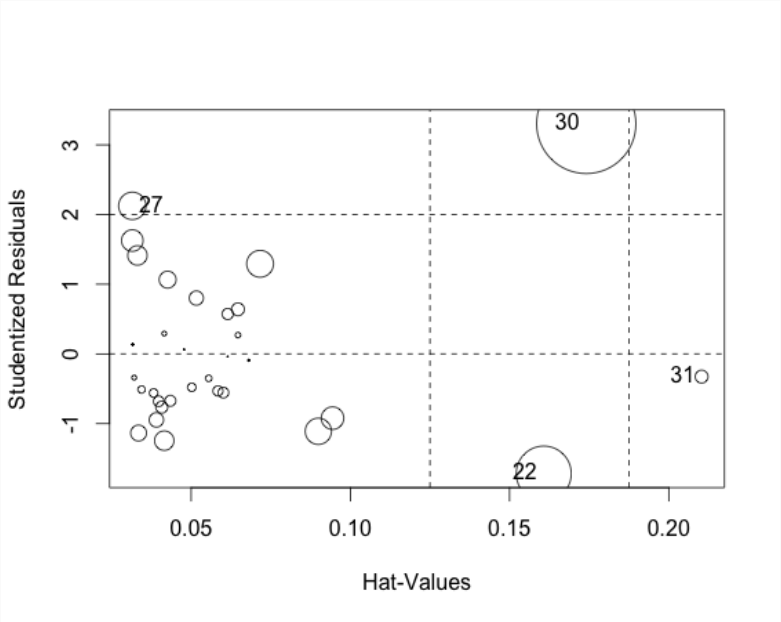
```
> outlierTest(lm1)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
30 3.302066        0.0025533    0.081704
```

The test also shows observation 30 to not be an outlier but we aren't done yet!

Leverage

- High Leverage: Observations that are relatively far from the center of the regression *space* may have greater influence on the regression model
- An observation's influence is a function of two factors: (1) leverage – how much the observation's x_i value differs from $\bar{x}$ and (2) Cook's Distance – the difference between $\hat{y}_i$ for the observation and y_i
- influenceIndexPlots and influencePlots are used to see the leverage and influence of each observation
- Focus on the second plot on the following slide: What you see is observation 30 flagged with the largest bubble, indicating that if this observation is removed from the sample the estimated slope will change appreciably because observation 30 is exerting a relatively large influence on the regression line


Diagnostic Plots
Cook's distance
Studentized residual
Bonferroni p-value
hat-values
Index
30
22
27
31


Studentized Residuals
Hat-Values
30
27
31
22

Refitting the Model sans Obs. #30

```
> lm1.no30 <- update(lm1, subset=-30)
> summary(lm1.no30)
Call:
lm(formula = age ~ proportion.black, data = LionNoses, subset = -30)
Residuals:
    Min      1Q  Median      3Q     Max
-2.0522 -0.9810 -0.4072 0.6353 3.4973
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.2938     0.5089   2.542   0.0166 *
proportion.black 8.8498     1.4175   6.243 8.19e-07 ***
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1        1
Residual standard error: 1.447 on 29 degrees of freedom
Multiple R-squared: 0.5734,  Adjusted R-squared: 0.5587
F-statistic: 38.98 on 1 and 29 DF, p-value: 8.191e-07

> compareCoefs(lm1, lm1.no30)
Call:
1:"lm(formula = age ~ proportion.black, data = LionNoses)"
2:c("lm(formula = age ~ proportion.black, data = LionNoses, ",
 "    subset = -30)")
               Est. 1  SE 1 Est. 2  SE 2
(Intercept)     0.879 0.569  1.294 0.509
proportion.black 10.647 1.510 8.850 1.418
```

Note how much both the intercept and the slope change once observation 30 is removed

Why is this observation problematic? See the data ...

Logit Models

Categorical Dependent Variables

Thus far y has been a numeric variable but what if $y = 1$ if a patient (or a species) survived and $y = 0$ if a patient (or the species) did not survive?

Let us make the example concrete by way of a specific dataset: SAheart[1]

Let $y = 1$ if the male has coronary heart disease and $y = 0$ otherwise

Let x be the male's age (in years)

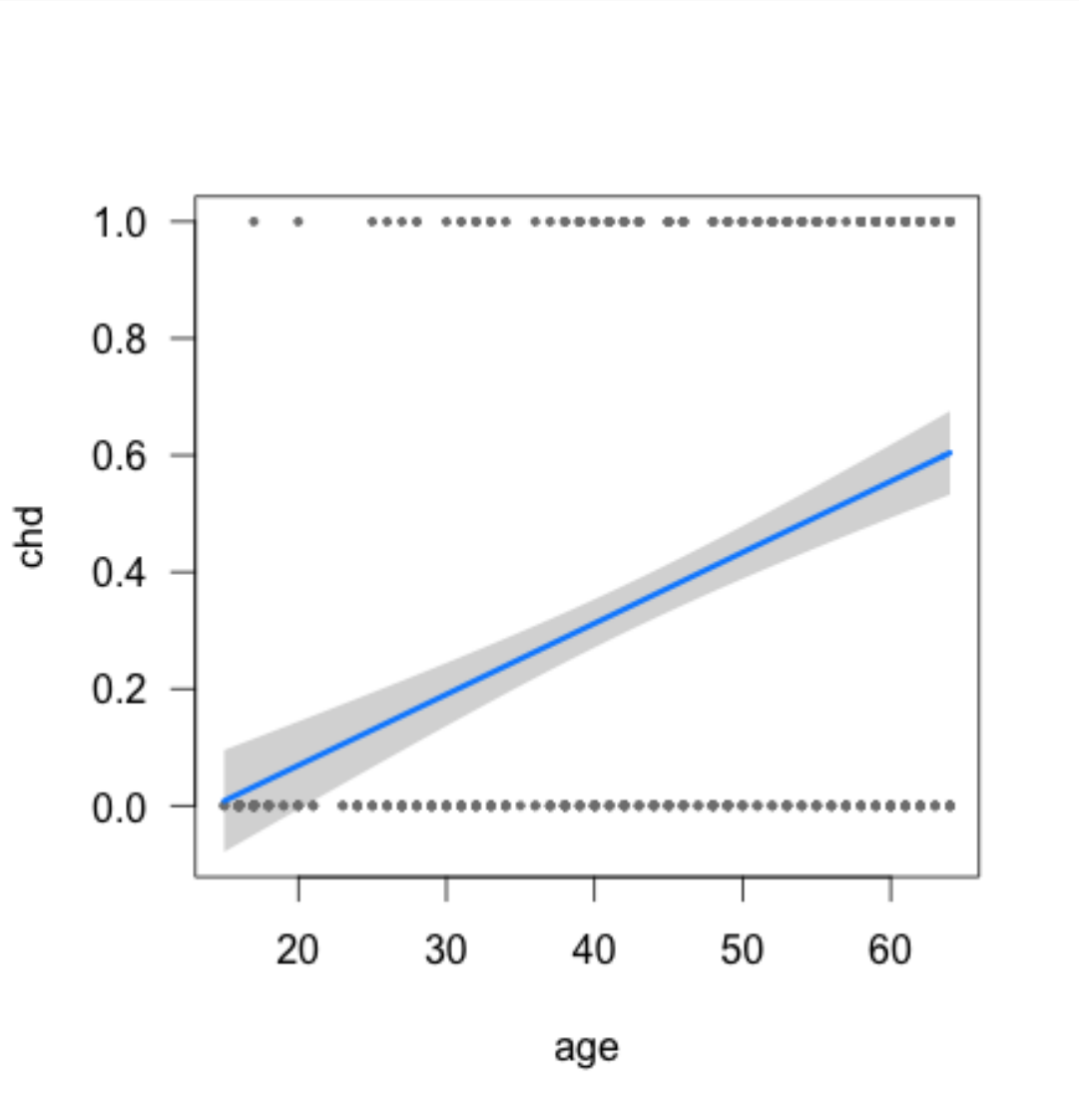
Question: Does a man's age influence the probability of coronary heart disease?

Let π_i be the probability that $y_i = 1$ and thus the probability that $y_i = 0$ must be $1 - \pi$

Maybe we can just fit a linear regression model $\pi_i = a + b(age)$?

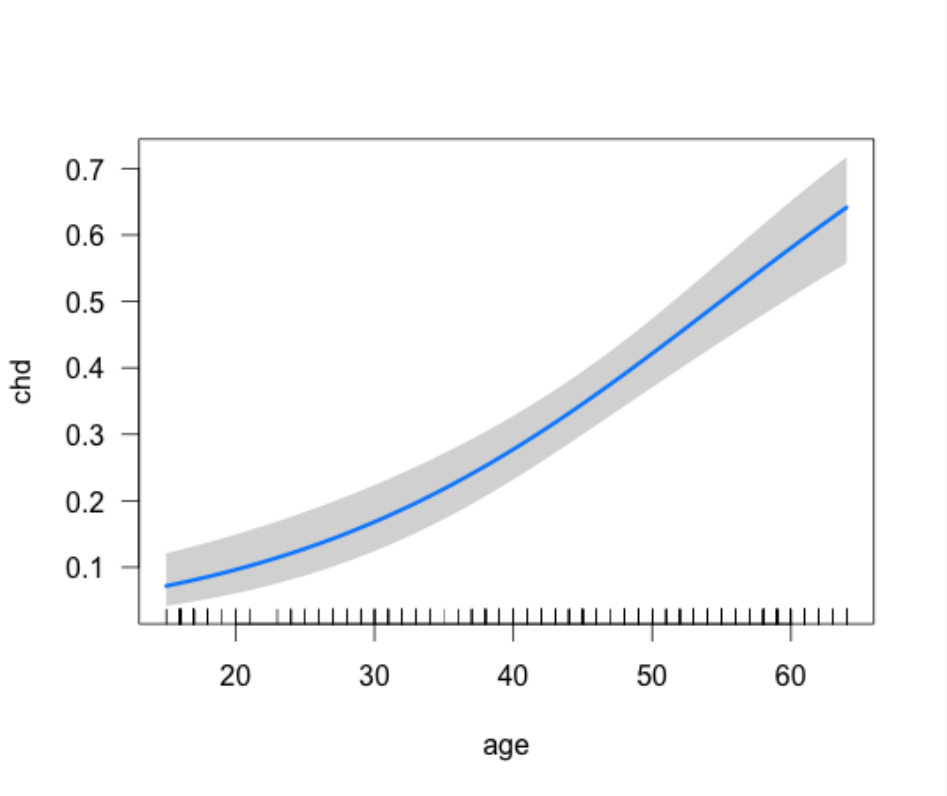
```
Call:
lm(formula = chd ~ age, data = SAheart)
Residuals:
    Min      1Q  Median      3Q     Max
-0.6039 -0.3729 -0.1418 0.4690 0.9676
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.17434  0.06380 -2.733 0.00653 **
age          0.01216   0.00141  8.621 < 2e-16 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
Residual standard error: 0.4424 on 460 degrees of freedom
Multiple R-squared: 0.1391,  Adjusted R-squared: 0.1372
F-statistic: 74.33 on 1 and 460 DF, p-value: < 2.2e-16
```

[1] These data are in the *ElemStatLearn* library

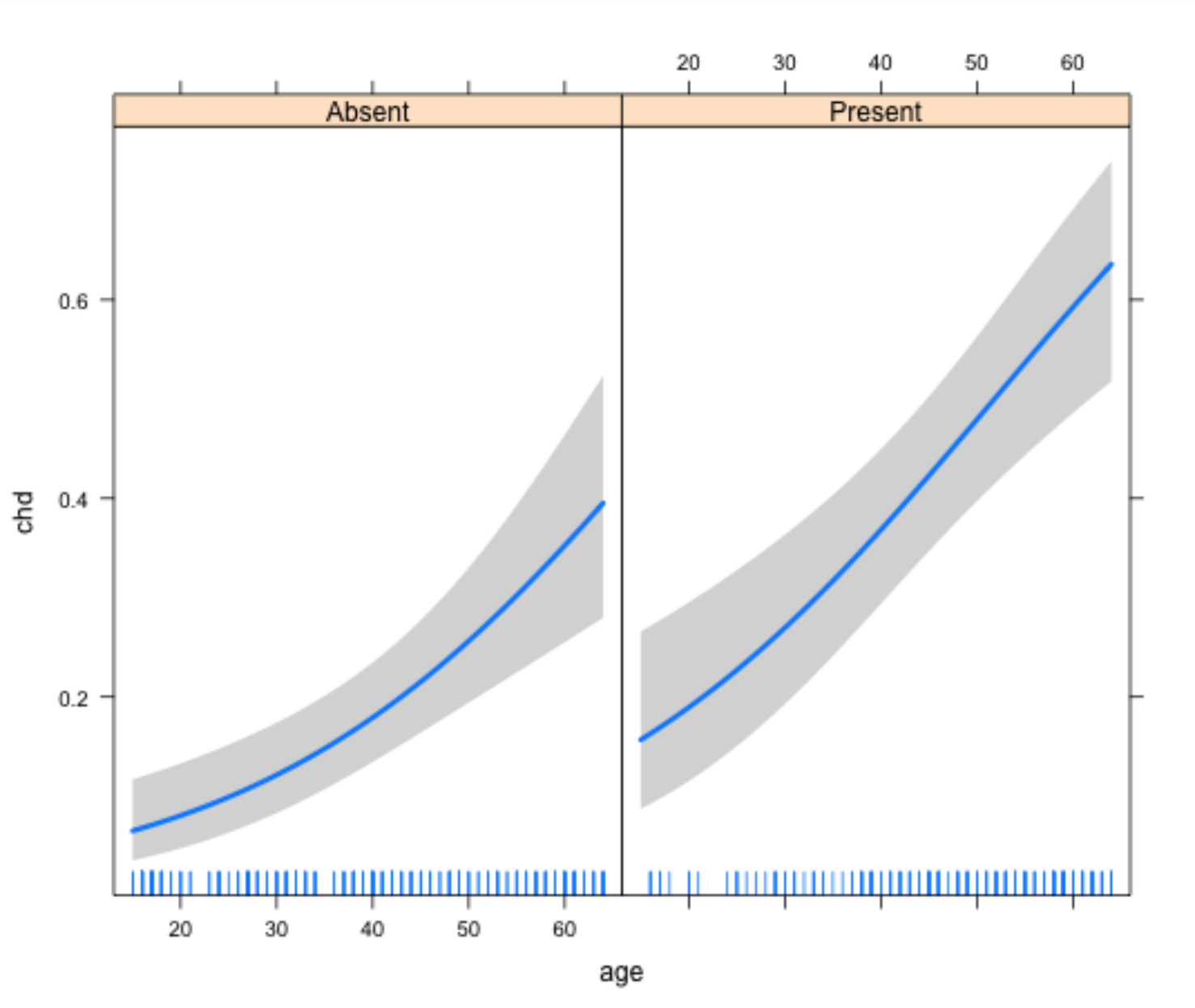


What are some of the things that strike you as odd?

```
Call:
glm(formula = chd ~ age, family = binomial(link = "logit"), data = SAheart)
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-1.4321 -0.9215 -0.5392  1.0952  2.2433
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.521710 0.416031 -8.465 < 2e-16 ***
age          0.064108  0.008532  7.513 5.76e-14 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
```



```
Call:
glm(formula = chd ~ age + sbp + tobacco + famhist, family = binomial(link = "logit"),
    data = SAheart)
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-1.8469 -0.8783 -0.4697  0.9668  2.4044
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.312010  0.783259 -5.505 3.69e-08 ***
age           0.045700  0.009861  4.634 3.58e-06 ***
sbp           0.005946  0.005497  1.082  0.27941
tobacco       0.082580  0.025821  3.198  0.00138 **
famhistPresent 0.982556  0.220512  4.456 8.36e-06 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
```


Absent
Present
20
30
40
50
60
chd
0.6
0.4
0.2
age