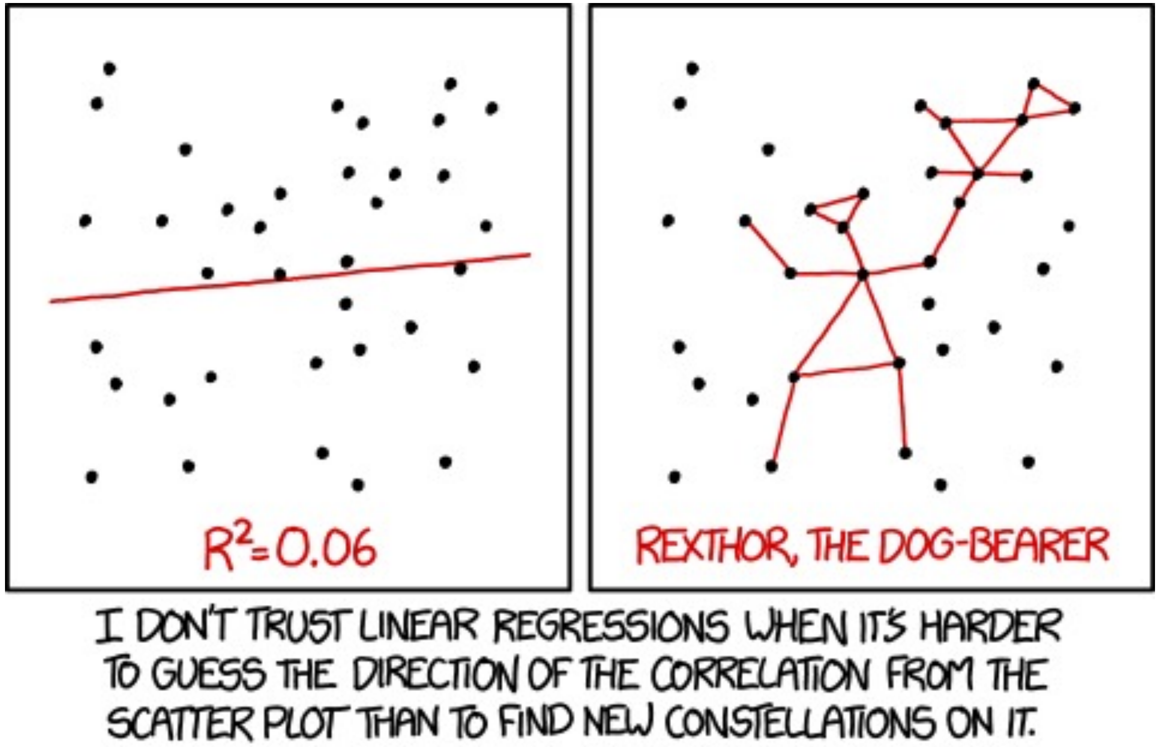
Multiple regression continued



Overview

Quick review of what we have covered in multiple regression

Log transformations of the response variable y

Multicollinearity

Polynomial regression

Quick review

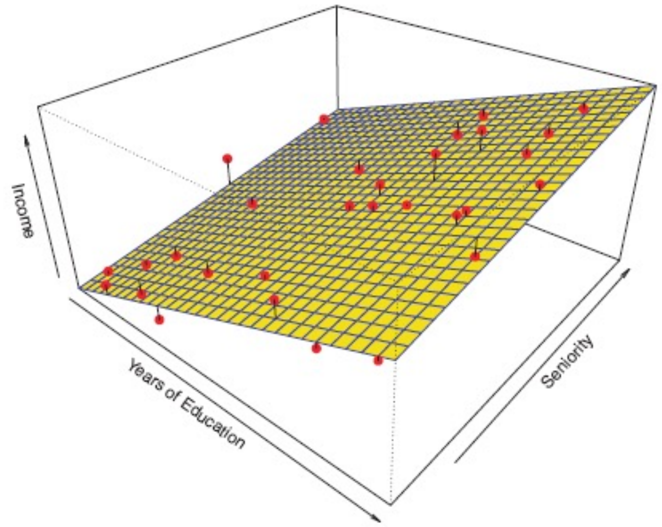
Multiple regression

In multiple regression we try to predict a quantitative response variable *y* using several predictor variables $x_1, x_2, \ldots, x_k$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \ldots + \hat{\beta}_k \cdot x_k$$

Goals:

- To make predictions as accurately as possible
- To understand which predictors (x) are related to the response variable (y)

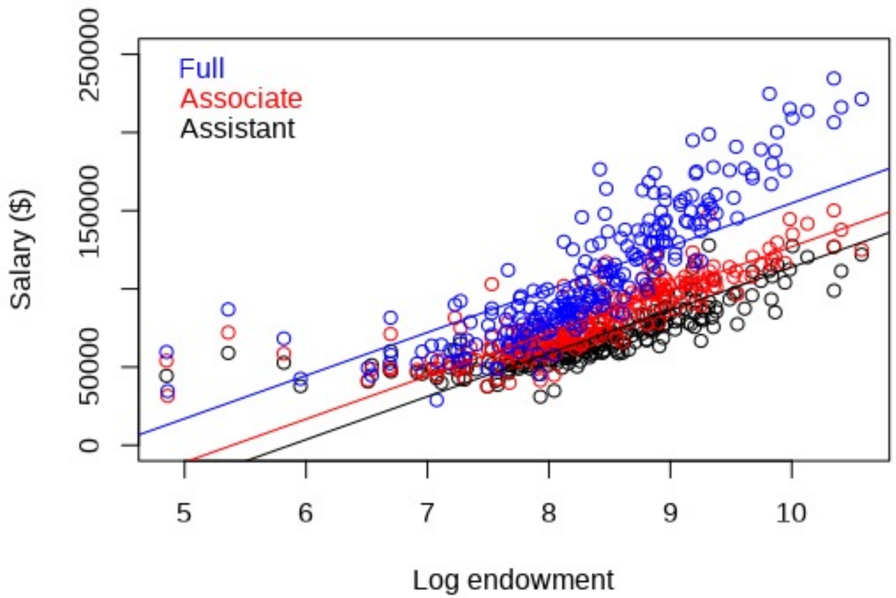


Categorical predictors

Predictors can be categorical as well as quantitative

- When a qualitative predictor has k levels, we need to use k -1 dummy variables to code it

Suppose we want to predict faculty salary *y* as a function of endowment x_1, with separate intercepts for faculty rank



$$x_{i1} = \log(\text{endowment})$$

$$x_{i2} = \begin{cases} 1 & \text{if assistant professor} \\ 0 & \text{otherwise} \end{cases}$$

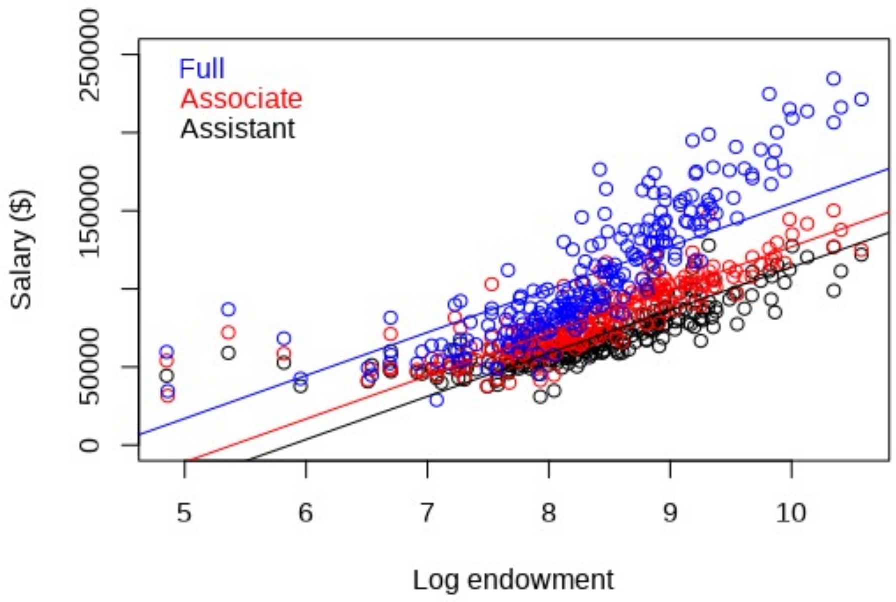
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 & \text{if assistant professor} \\ \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_3 & \text{if associate professor} \\ \hat{\beta}_0 + \hat{\beta}_1 x_{i1} & \text{if full professor} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{if associate professor} \\ 0 & \text{otherwise} \end{cases}$$

Categorical predictors

Predictors can be categorical as well as quantitative

- When a qualitative predictor has k levels, we need to use k -1 dummy variables to code it



Suppose we want to predict faculty salary as a function of endowment with separate intercepts for faculty rank

```
> summary(fit_prof_rank_offset)

Call:
lm(formula = salary_tot ~ log_endowment + rank_name, data = IPED_2)

Residuals:
   Min     1Q Median     3Q    Max
-52464 -10844  -2703   6936  74994

Coefficients:
                    Estimate Std. Error t value            Pr(>|t|)
(Intercept)        -120822.1     6713.9  -18.00 <0.0000000000000002 ***
log_endowment        27569.9      791.7   34.82 <0.0000000000000002 ***
rank_nameAssociate  -27855.4     1685.5  -16.53 <0.0000000000000002 ***
rank_nameAssistant  -40973.7     1685.5  -24.31 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18370 on 707 degrees of freedom
Multiple R-squared:  0.7192,	Adjusted R-squared:  0.718
F-statistic: 603.7 on 3 and 707 DF,  p-value: < 0.00000000000000022
```

$$\hat{y}_i = \begin{cases} \boxed{\hat{\beta}_0} + \boxed{\hat{\beta}_1} x_{i1} + \boxed{\hat{\beta}_2} & \text{if assistant professor} \\ \boxed{\hat{\beta}_0} + \boxed{\hat{\beta}_1} x_{i1} + \boxed{\hat{\beta}_3} & \text{if associate professor} \\ \boxed{\hat{\beta}_0} + \boxed{\hat{\beta}_1} x_{i1} & \text{if full professor} \end{cases}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$$

$$= -120,822 + 27,570 x_{i1} - 40,973 x_{i2} - 27,855 x_{i3}$$

Interaction terms

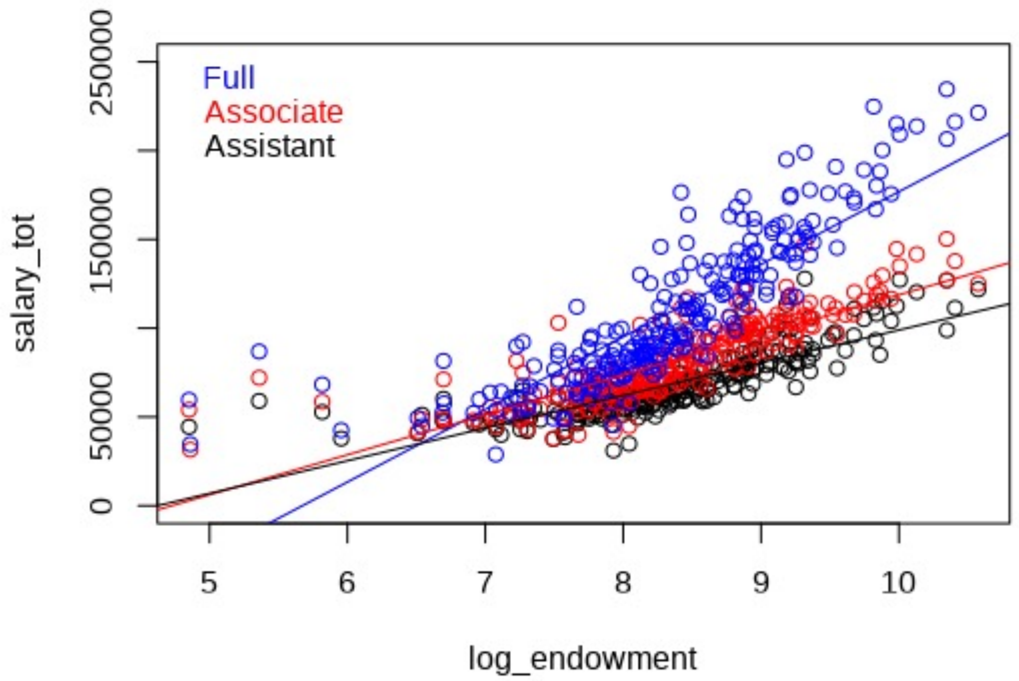
An ***interaction effect*** occurs when the response variable y is influenced by the levels of two or more predictors in a non-additive way

We can model this using an equation with an interaction term

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_3 (x_1 \cdot x_2) + \epsilon$$

An interaction term between a quantitative and categorical variable corresponds to different slopes depending for the quantitative variable depending on the value of the categorical variable

Interaction terms



If Full Professor:

$$\text{salary} \approx \beta_0 + \beta_1 \cdot \text{endowment}$$

If Assistant Professor:

$$\text{salary} \approx (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \text{endowment}$$

Additive term if Assistant Professor

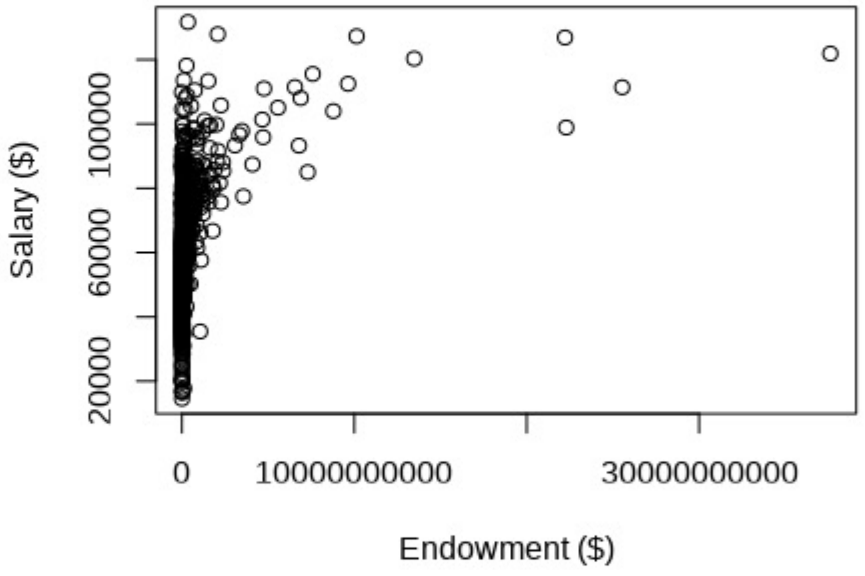
Change in slope if Assistant Professor

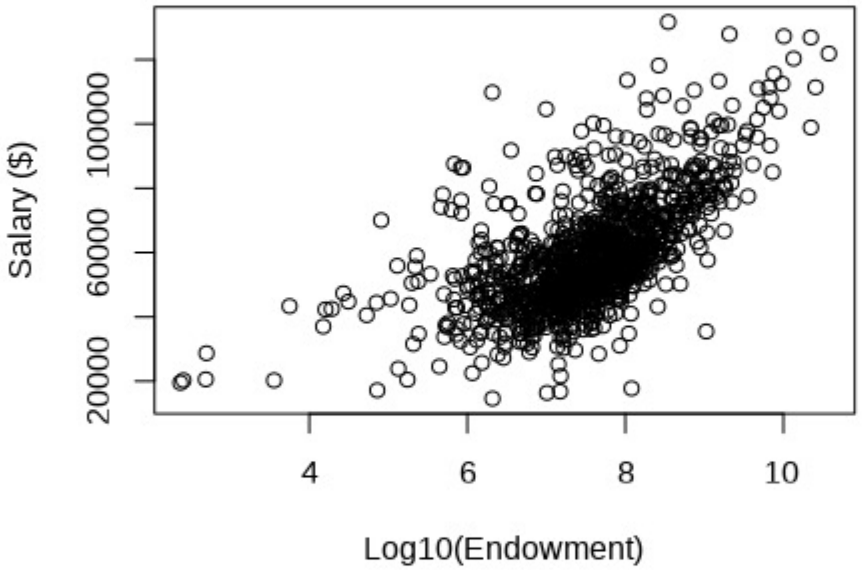
$$x_{i2} = \begin{cases} 1 & \text{if assistant professor} \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i1} \cdot x_{i2}$$

Log transformation of the response variable y

As we've seen, we can take a log transformation of an *explanatory x* variable to make a non-linear relationship more linear

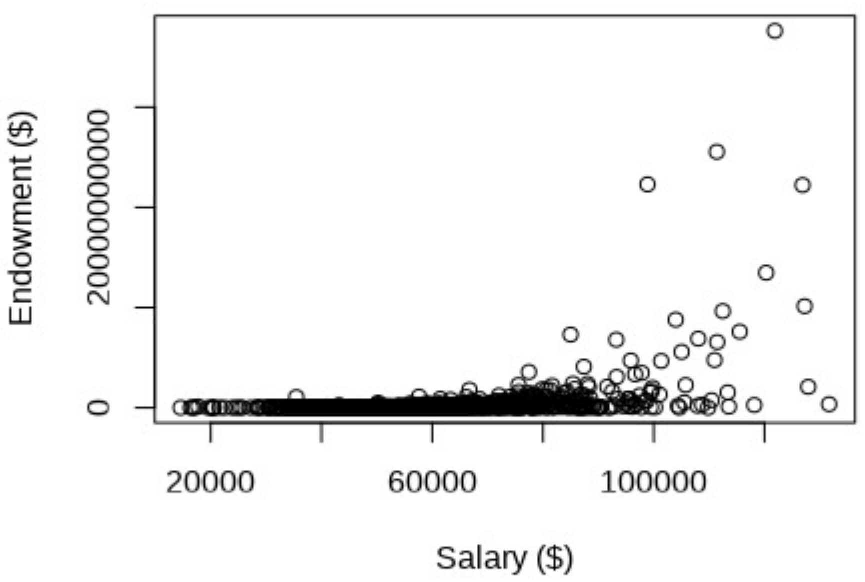


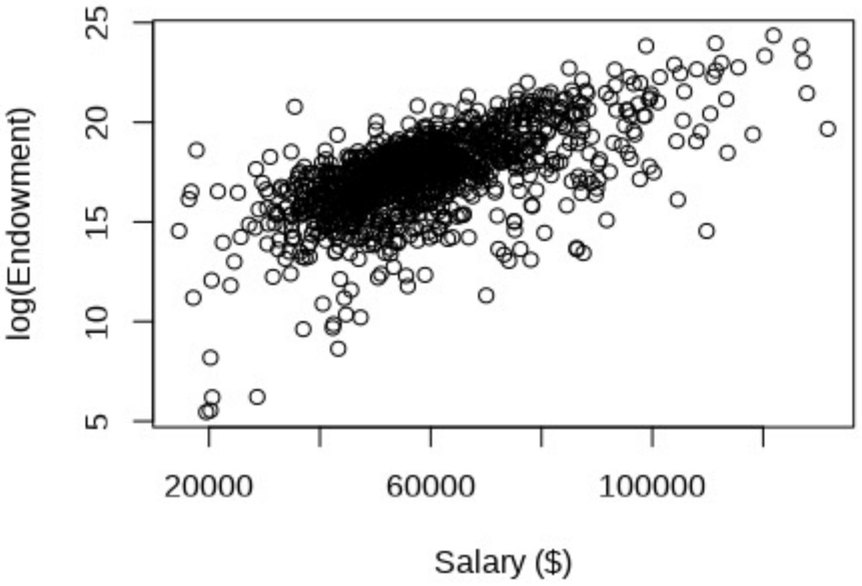


Log transformation of the response variable y

Often, it can be useful to take log transformation of a *response variable y* to make the relationship more linear

- This can also be useful to deal with heteroskedasticity





Log transformation of the response variable y

How can we interpret the regression coefficients when we have taken a log transformation of the response variable y?

$$log(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

If we exponentiate both sides we get:

$$y = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1} = e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_1 x_1}$$

If we increase x by 1, we multiply the previous predicted value of ŷ by $e^{\hat{\beta}_1}$

$$\hat{y} = \hat{f}(x+1) = e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_1 x_1 + 1} = e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_1 x_1} \cdot e^{\hat{\beta}_1} = \hat{f}(x) \cdot e^{\hat{\beta}_1}$$

Log transformation of the response variable y

Side note: Often the natural (base e) log of y is used because for small values of $\hat{\beta}$

$$e^{\hat{\beta}} \approx 1 + \hat{\beta}$$

This is used as a justification for using the natural log, since this allows one to directly see what $e^{\hat{\beta}}$ approximately is from just looking at $\hat{\beta}$

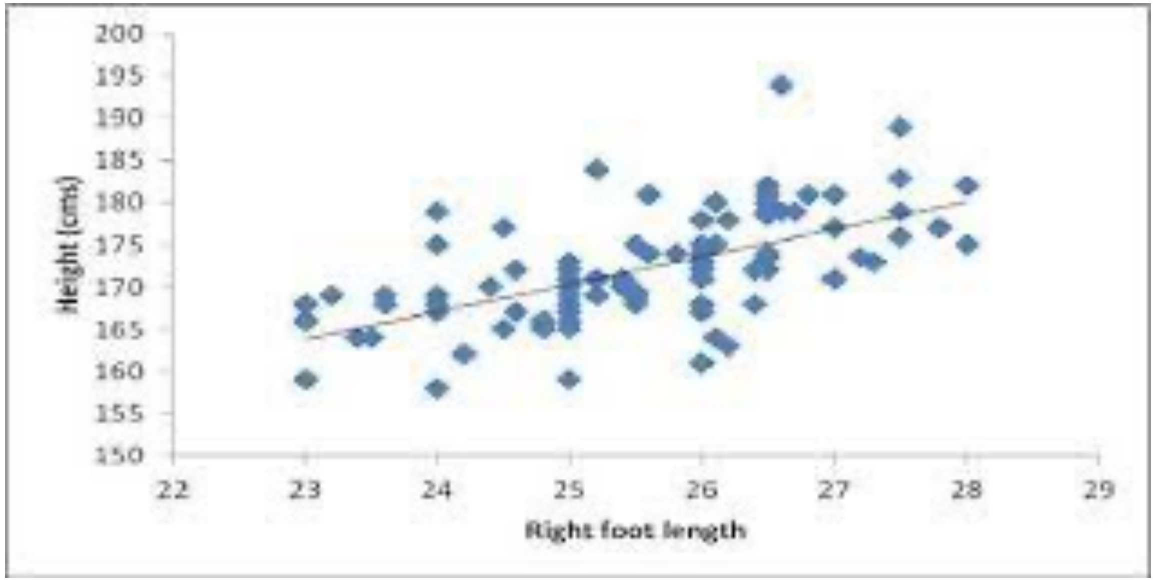
- Although it's not very hard to use the exp() on the regression coefficients in R

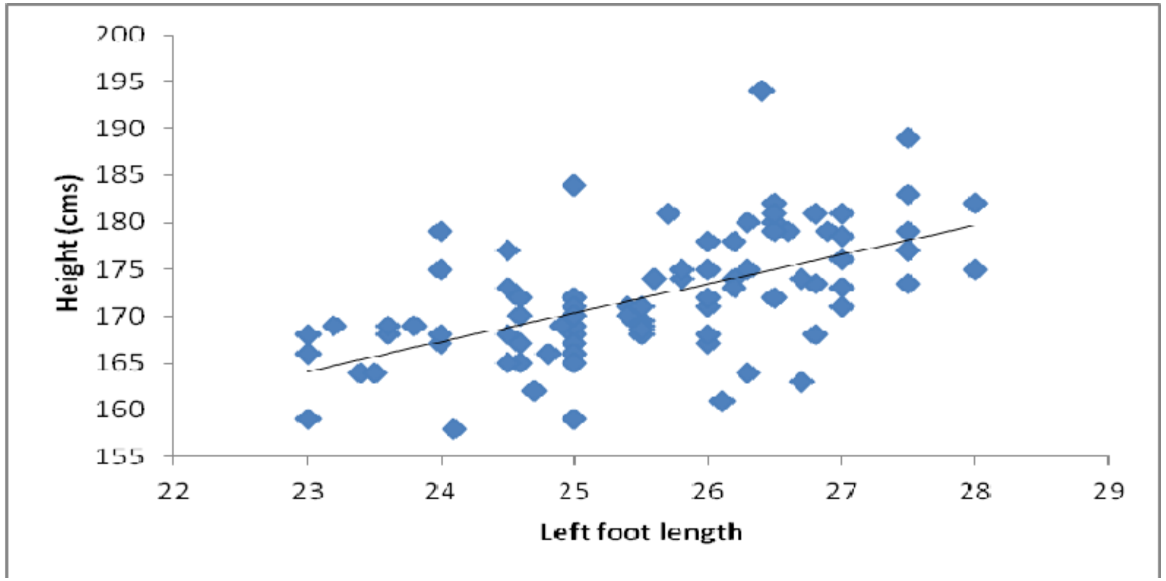
Let's try it in R…

Multicollinearity

Multicollinearity occurs when two or more variables are closely related to each other

- E.g., if they have a high correlation





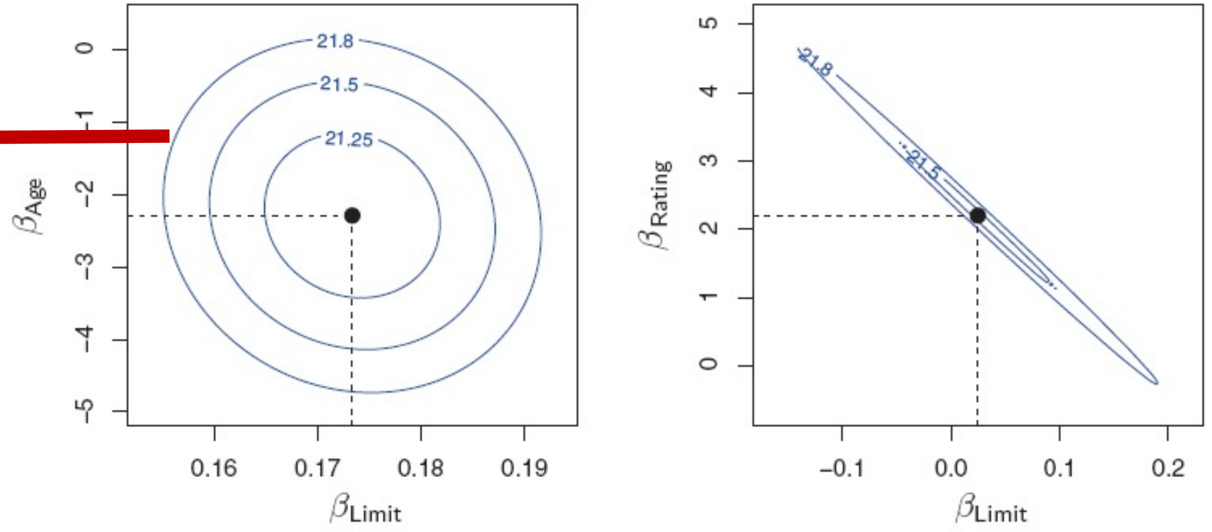
Multicollinearity

Multicollinearity can make our estimate of the regression coefficients unstable

- i.e., a large range of coefficient β-hat values give the same SSResidual and $\hat{\sigma}_\varepsilon$

Contours of equal

$\hat{\sigma}_\varepsilon$ value



This increases our estimate of the variance of the coefficients we measure and hence can decrease the power to detect a statistically significant predictor

Multicollinearity

The **variance inflated factor** is a statistic that can be computed to test for multicollinearity

$$VIF_i = \frac{1}{1-R_i^2}$$

where R_i^2 is the coefficient of multiple determination for a model to predict x_i using the other predictors in the model

<u>Rule of thumb</u>: suspect multicollinearity for VIF > 5

car::vif(lm_fit)

Are any of the predictors x_i related to y?

We can set this up as a hypothesis test:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_A: At least one $\beta_j \neq 0$

We can run a parametric hypothesis test based on an F statistic to test this hypothesis

```
Call:
lm(formula = R ~ X1B + X2B + X3B + HR + BB + X1Bn + X2Bn + X3Bn +
    XHRn + XBBn, data = team_batting2)

Residuals:
    Min      1Q  Median      3Q     Max
-78.695 -15.457  -0.798  15.480  76.092

Coefficients:
               Estimate  Std. Error t value             Pr(>|t|)
(Intercept)  -574.88241    20.89696 -27.510 <0.0000000000000002 ***
X1B            -0.08976     0.43995  -0.204                0.838
X2B             1.70203     1.36050   1.251                0.211
X3B            -0.20163     4.71591  -0.043                0.966
HR              1.19258     1.47183   0.810                0.418
BB              0.24157     0.65658   0.368                0.713
X1Bn         3930.66847  2443.75215   1.608                0.108
X2Bn        -4839.59898  7517.51009  -0.644                0.520
X3Bn         8493.67060 26119.44048   0.325                0.745
XHRn         2061.44301  8146.72963   0.253                0.800
XBBn          588.32226  3628.53349   0.162                0.871
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.61 on 1140 degrees of freedom
Multiple R-squared:  0.9297,	Adjusted R-squared:  0.929
F-statistic:  1507 on 10 and 1140 DF,  p-value: < 0.00000000000000022
```

summary(lm_fit)

None of the coefficients are significant at the $\alpha = 0.05$ level

Overall H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ is highly significant

This can happen when there is multicolinearity

Polynomial regression

Polynomial regression extends linear regression to non-linear relationships by including nonlinear transformations of predictors

$$\begin{aligned} \text{salary} = \; & \beta_0 + \beta_1 \cdot \text{endowment} \\ & + \beta_2 \cdot (\text{endowment})^2 + \\ & + \beta_3 \cdot (\text{endowment})^3 + \varepsilon \end{aligned}$$

Still a linear equation but non-linear in original predictors

Polynomial regression

Polynomial regression extends linear regression to non-linear relationships by including nonlinear transformations of covariates

We can compare model fits by:

- Assessing if higher order terms are statistically significant
- Looking at the r^2 values
- Running hypothesis tests comparing nested models
- Etc.

Let’s try it in R…