7: Linear Regression Models with a Single Predictor (Part 2)

Packages Needed To Recreate Code on Slides

```
library(mfp)        # for bodyfat data set
library(MASS)       # for cats data set
library(tidyverse) # for ggplot functions/plotting
data(bodyfat)       # load data
```

Warning: It is not expected that you understand all the R code in this presentation right now. You will go over more R code in SDS 100 in the coming weeks. However, you are welcome to try to make these plots on your own.

Linear Regression with a Single Predictor

What do we mean by saying *"linear regression model with a single predictor"*

- **predict**: indicate in advance
  - x *can help us indicate what* y *will be.*
- **regress**: to tend to approach or revert to a value/relation
  - x *&* y *values approach a common relationship.*
- **linear**: $y = b_0 + b_1 x$
  - x *&* y *relationship can roughly be described by a straight line.*
- **model**: an informative representation of an object, person or system.
  - *Educated guess for* b_0 *&* b_1 *describing the relationsip of* x *&* y*.*

Linear Regression with a Single Predictor

Linear regression models can be used for:

1) prediction

If I have a new data value x^*, can I guess what it's corresponding value for y would be?

2) evaluate whether there is a linear relationship between two numerical variables.

Does a linear relationship exist between x and y?

Does this fashion hack work?

*Image from TikTok @nicolefay_*

The Latest TikTok Hack To Fitting Jeans Without Trying Them On

BY MIA UZZELL POSTED ON AUGUST 24, 2022

Evading the dreaded fitting room just got a lil' easier.

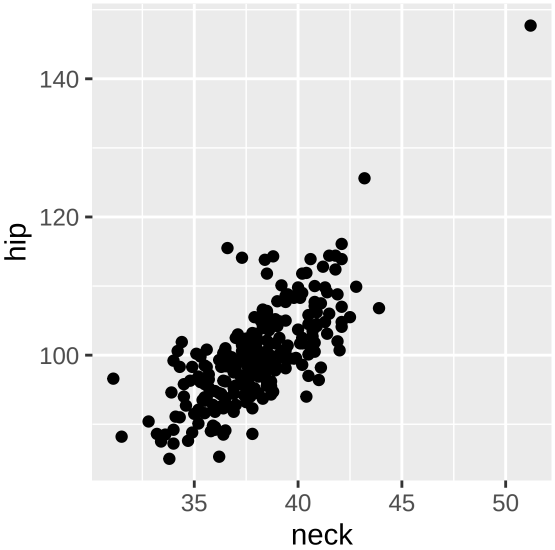
TikTok's newest craze has cropped up in its whirlpool of fashion trends. And it has set a pretty lofty expectation: selecting the perfect pair of denim bottoms sans the anxiety of the fitting room.

Linear Regression with a Single Predictor

- Want to see if the circumference of our hips (`hip`) is related to the circumference of our neck (`neck`).
- If so, we can avoid dressing rooms!
- Data were supplied by Dr. A. Garth Fisher, Human Performance Research Center, Brigham Young University, who gave permission to freely distribute the data and use them for non-commercial purposes.
- Data set is from 252 men, and records various body measurements.

Linear Regression with a Single Predictor

```
ggplot(bodyfat, aes(x = neck, y = hip))+
  geom_point()
```



Note on Notation

The regression model assumes that *true* relationship is the following:

$$Y = \beta_0 + \beta_1 X$$

However, in practice there are nuances we can not capture. So what we actually observe is

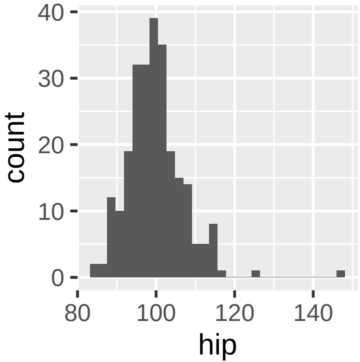
$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Note on Notation

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

That is, y_i is variable, and that variability can be broken up in two parts:

- variability that can be explained by neck size
- everything else, the 'left over'



Note on Notation

β_0 and β_1 are considered to be the **unknown** *truth.*

We want to estimate them!

- We denote estimates for β_0 as:

$$\hat{\beta}_0 \text{ or } b_0$$

- We denote estimates for β_1 as:

$$\hat{\beta}_1 \text{ or } b_1$$

Fitting a Linear Regression Model

Want to find b_0 and b_1 such that:

$$min \left\{ \sum_{i=1}^{n} [e_i]^2 \right\}$$

Equivalently:

$$min \left\{ \sum_{i=1}^{n} [y_i - (b_0 - b_1 x_i)]^2 \right\}$$

Fitting a Linear Regression Model

We can then use techniques from calculus to identify these values

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} \left[y_i - (b_0 - b_1 x_i)\right]^2 \overset{set}{=} 0$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^{n} \left[y_i - (b_0 - b_1 x_i)\right]^2 \overset{set}{=} 0$$

These solutions can be written as functions of the summary statistics we have already seen:

- $b_1 = r \frac{s_y}{s_x}$
- $b_0 = \overline{y} - b_1 \overline{x}$

Exercise

Can you find the estimates for b_0 and b_1 with the following summary statistics?

```
c(mean(bodyfat$hip), sd(bodyfat$hip))
```

```
[1] 99.904762  7.164058
```

```
c(mean(bodyfat$neck), sd(bodyfat$neck))
```

```
[1] 37.992063  2.430913
```

```
cor(bodyfat$neck, bodyfat$hip)
```

```
[1] 0.7349579
```

Interpretation

What do b_0 and b_1 really tells us?

- The expected mean value for Y when $X = 0$ is b_0
- For every one unit of increase in X we expect the mean value of Y to change by b_1 units

This wording is important!

Try this on your own for this example.

Interpretation

Example:

- The average value for `hips` is 17.615 when `neck` is 0.
- For every one cm of increase in `neck` the expected mean value of `hips` to increase by 2.16 cm

Warning:

- The coefficient b_0 dose not always have a useful interpretation.
- X units and Y units can be different.

Evaluating the model fit

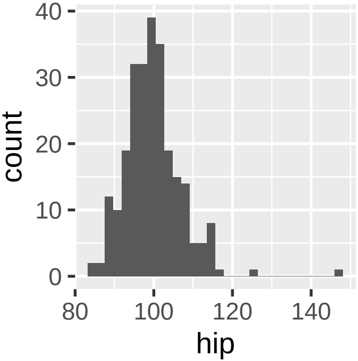
- Previously we used r as a quick and simple gauge for assessing the relationship.
- To use a more rigorous evaluation method, we need more tools.

Sum of Squares

We can measure the variability of the Y values by how far they tend to fall from their mean $\overline{y}$. This is called **total sum of squares (SST)**.

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- Similar to variance.
- Describes overall variability.



Sum of Squares

Recall our start! The variation in Y can be explained by two parts,

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Through algebraic manipulation we can rewrite SST,

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

We denote the above components as

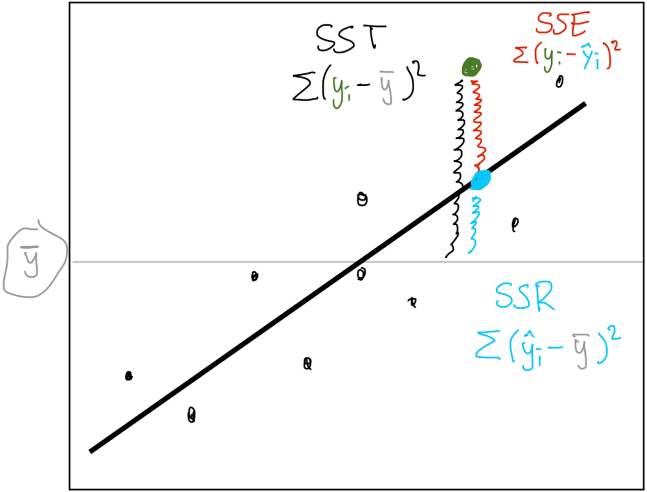
$$SST = SSR + SSE$$

Sum of Squares

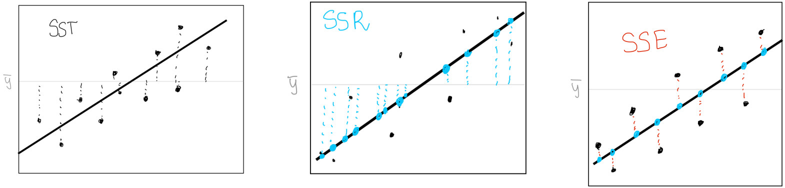
$$SST = SSR + SSE$$

- total sum of squares (SST): the total variability of Y
- regression sum of squares (SSR): the variability of Y explained by the model (X)
- error sum of squares (SSE): the variability of Y NOT explained by the model. What is 'left-over'

Sum of Squares



Sum of Squares



Coefficient of Determination

- **Coefficient of Determination** (R^2): measures the proportion of the variation in the outcome variable Y that is explained by the linear regression model with predictor X

$$R^2 = \frac{SST - SSE}{SST} = \frac{SSR}{SST}$$

- Note: R^2 is just the correlation squared!

Linear Regression Models in R with Body Data

Estimates for b_0 and b_1

```
fit <- lm(hip ~ neck, data = bodyfat)
summarize_fit <- summary(fit)
summarize_fit$coefficients
```

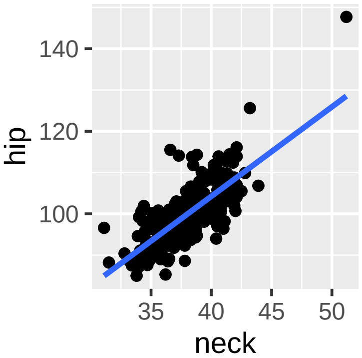
```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 17.615163  4.8116950  3.660906 3.066444e-04
neck         2.165968  0.1263926 17.136832 4.574017e-44
```

Linear Regression Models in R with Body Data

Linear regression model

$$\widehat{\texttt{hip}} = 17.62 + 2.17\ \texttt{neck}$$

```
ggplot(bodyfat, aes(x = neck, y = hip))+
  geom_point()+
  geom_smooth(method = "lm", se = F)
```



Linear Regression Models in R with Body Data

Obtaining R^2

```
summarize_fit$r.squared
```

```
[1] 0.5401631
```

Interpreting R^2

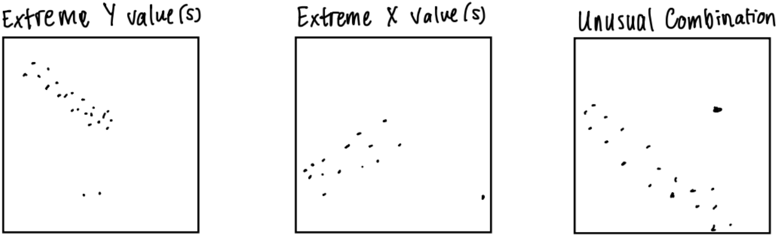
About 54% of the variability of `hip` can be accounted for by the model (the `neck` variable)

Outliers in Regression

There are many types of outliers in regression models

- extreme X values
- extreme Y values
- extreme/unusual combinations of X and Y

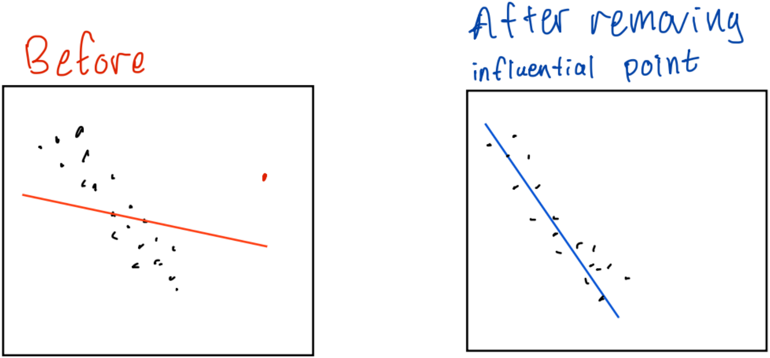
Outliers in Regression



Outliers in Regression

- We worry about outliers when cause undue influence and *pull* the line away from the cloud of points.
- If we had fitted a line without a point and it would be dramatically different we call this point an **influential point**.
- Do not hastily remove outliers, they could be the most important!

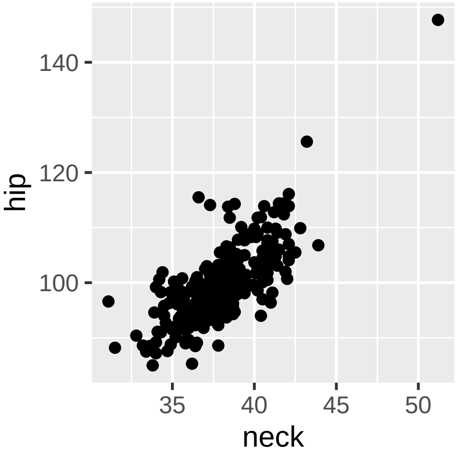
Influential Points



If we see a big change in the linear regression line after removing a value we call this point **influential**.

Extrapolation

What if a women wanted to check if this fashion hack worked? Or a child? Could they use the model we made?



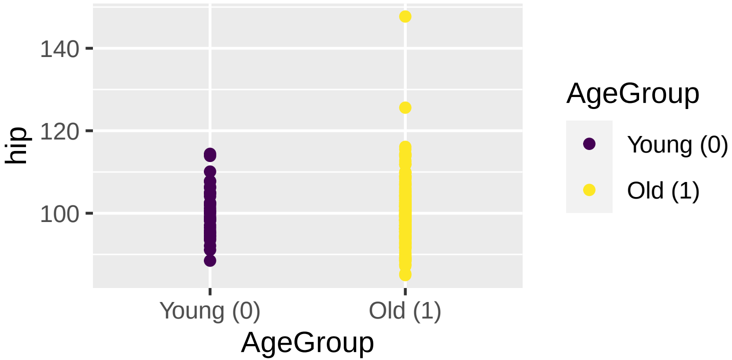
Categorical Variable with Two Levels

- Sometimes we wish to use a categorical predictor variable.
- When we only have two levels we can code them with an **indicator** variable
- We use 0 for one category and 1 for the other category.

Note:

- It does not matter which category is 0 or 1.
- We can even do a 1 and 2 coding, or 3 and 4. If we do this though our intrepretation changes.

Categorical Variable with Two Levels



Categorical Variable with Two Levels

Interpretation for b_0, b_1 with an indicator variable

- *Interpret the intercept:* The expected mean value of Y for a subject in the level-0 group is b_0
- *Interpret the slope*: The expected mean value of Y changes by b_1 units when a subject is in level-1 group in comparison to the level-0 group

Why say *expected mean*?

Categorical Variable with Two Levels

Interpretation for b_0, b_1 with an indicator variable

- *Interpret the intercept:* The expected mean value of Y for a subject in the level-0 group is b_0
- *Interpret the slope*: The expected mean value of Y changes by b_1 units when a subject is in level-1 group in comparison to the level-0 group

Why say *expected mean*?

This is the mean of the data, but the model is an estimate of the relationship. We do not know the mean of the population.

Categorical Variable with Two Levels

Suppose $\widehat{\texttt{hip}} = 100.15 - 0.48\ \texttt{AgeGroupOld}$

Example:

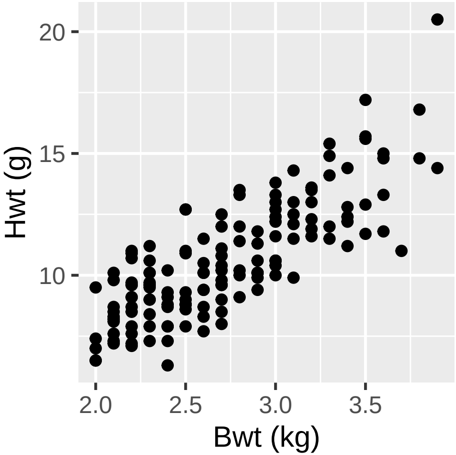
- *Interpret the intercept:* The expected mean value of `hip` for a subject in the `Young (0)` group is 100.15
- *Interpret the slope*: The expected mean value of `hip` decreases by 0.4842 cm when a subject is in `Old (1)` group in comparison to the `Young (0)` group

Example Problems

Cats Data

Recall the cats data set:

```
ggplot(cats, aes(x = Bwt, y = Hwt))+
  geom_point() + labs(x = "Bwt (kg)", y = "Hwt (g)")
```



Example 1

Suppose:

$$\overline{Bwt} = 2.724 \qquad \overline{Hwt} = 10.63$$

$$s^2_{Bwt} = 0.235 \qquad s^2_{Hwt} = 5.93 \qquad R^2 = 0.65$$

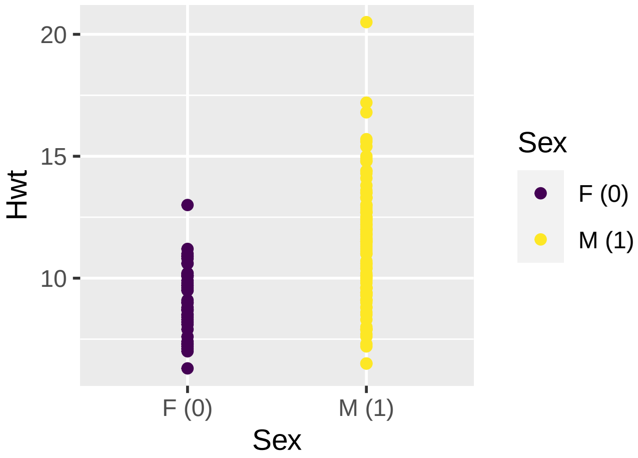
Answer the following prompts:

1. Calculate r
2. Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$
3. Write out the linear regression model.
4. Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$ in context.

Example 1- Answers

1. 0.80
2. -0.3567, 4.0341
3. $\widehat{Hwt} = -.3567 + 4.0341\ \texttt{Bwt}$
4. Interpretation
   - The expected mean value for Hwt when Bwt $= 0$ is $-.3567$.
   - For every one kg of increase in Bwt we expect the mean value of Y to increase by 4.0341 g.

Example 2



Example 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.262404	0.1980372	51.820576	3.984219e-94
SexM	1.499457	0.2800670	5.353924	3.379786e-07

1. Write out the linear regression model.
2. Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$ in context.
3. If my cat is a female that has a heart weight of 12 g, what is her corresponding residual?

Example 2 - Solutions

1. $\widehat{\texttt{Hwt}} = 10.26 + 1.50\ \texttt{SexM}$
2. Interpretation
   - The mean value of `Hwt` when `Sex = F` is 10.26
   - We expect the mean value of `Hwt` to increases by 1.50 grams when `Sex = M` in comparison to the `Sex = F` group
3. -1.74