# SDS358: Applied Regression Analysis

## Day 8: SLR: Inference

Dr. Michael J. Mahometa

## Agenda for Today:

- Simple Linear Regression model residuals
  - ANOVA table
- "Test" for slope significance

# Research Queastion:

Can perceived social support significantly predict life satisfaction in unemployed Spanish adults?



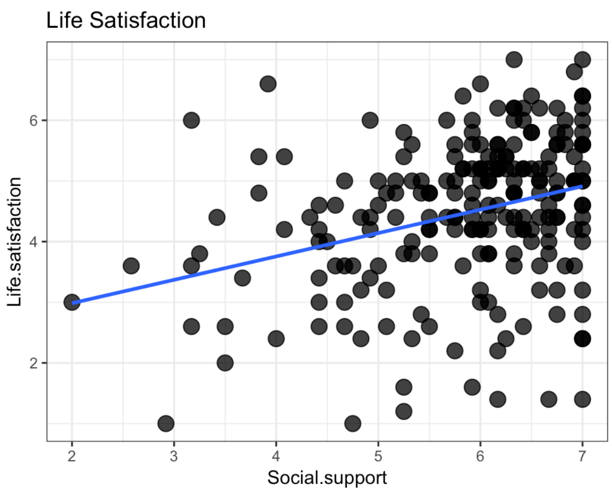
# The basics: correlation

```
vars <- c("Life.satisfaction", "Social.support")
cor(select(unemp, one_of(vars)), use="pairwise.complete.obs")
```

```
##                   Life.satisfaction Social.support
## Life.satisfaction         1.0000000      0.3146274
## Social.support            0.3146274      1.0000000
```

# The basics: visual

```
s <- simpleScatter(unemp, Social.support, Life.satisfaction,
                   title="Life Satisfaction", line=TRUE)
s
```



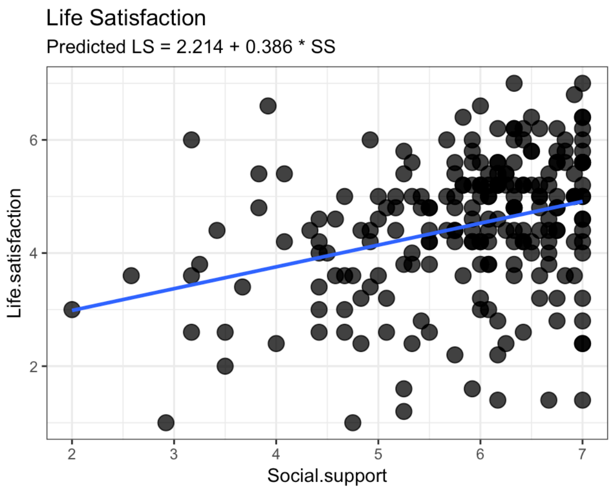



# The basics: SLR model

```
ls_mod <- lm(Life.satisfaction ~ Social.support, unemp)
ls_mod
```

```
##
## Call:
## lm(formula = Life.satisfaction ~ Social.support, data = unemp)
##
## Coefficients:
##    (Intercept)  Social.support
##         2.2136          0.3855
```

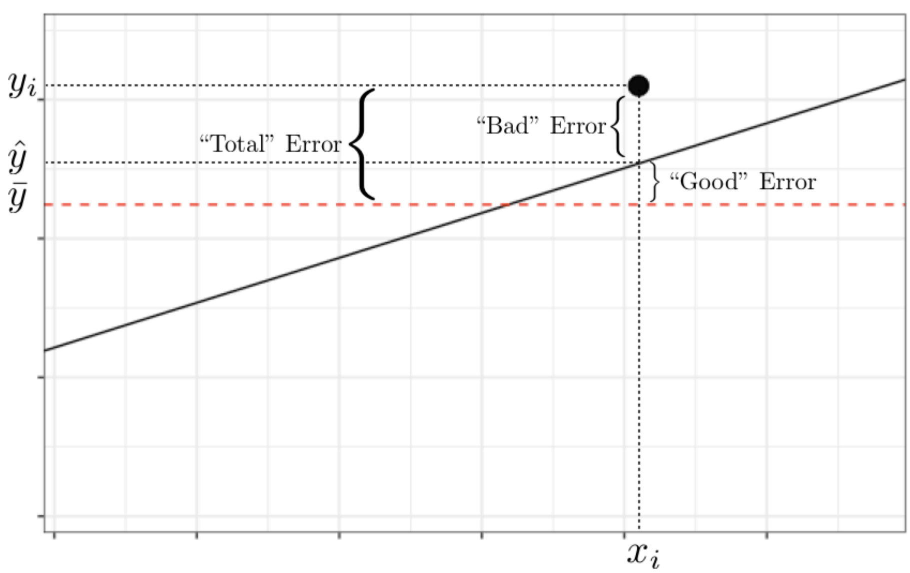
# The basics: visual






# Three kinds of varaince

## The basics



# Three kinds of varaince: Total

- First error: the "*Total*" error (without knowing anything else)

$$\Sigma(y_i - \bar{y}_i)^2$$

```
unemp$dev <- unemp$Life.satisfaction - mean(unemp$Life.satisfaction, na.rm=TRUE)
unemp$dev_sq <- unemp$dev^2
dev_sq <- sum(unemp$dev_sq, na.rm = TRUE)
dev_sq
```

```
## [1] 334.2793
```



# Three kinds of varaince: Model

- Second error: the "*Model*" variance (How much you improve, by using a model)

$$\Sigma(\hat{y}_i - \bar{y}_i)^2$$

```
unemp$pred <- predictValues(ls_mod)
unemp$mod <- unemp$pred - mean(unemp$Life.satisfaction, na.rm=TRUE)
unemp$mod_sq <- unemp$mod^2
mod_sq <- sum(unemp$mod_sq, na.rm = TRUE)
mod_sq
```

```
## [1] 32.98348
```

## Three kinds of varaince: Error

- Third error: the "*Error*" variance (How much are you still missing, even with the model?)

$$\Sigma(y_i - \hat{y}_i)^2$$

```
unemp$res <- unemp$Life.satisfaction - unemp$pred
unemp$res_sq <- unemp$res^2
res_sq <- sum(unemp$res_sq, na.rm = TRUE)
res_sq
```

```
## [1] 300.1683
```



## Three kinds of varaince

### All together now

```
dev_sq
```

```
## [1] 334.2793
```

```
mod_sq
```

```
## [1] 32.98348
```

```
res_sq
```

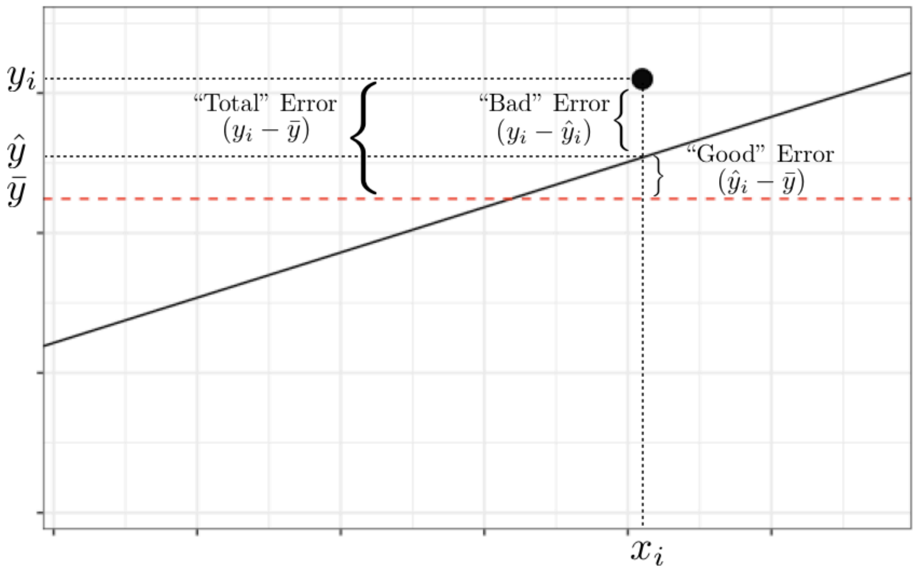
```
## [1] 300.1683
```

```
mod_sq + res_sq
```

```
## [1] 333.1518
```

# Three kinds of varaince

All together now






# Three kinds of variance

- With these three types of error, does it remind us of anything?
- How about an ANOVA table?

| Source | Sums of Squares | df | Mean Squares | F-value |
|---|---|---|---|---|
| **Regression** | $\Sigma(\hat{y}_i - \bar{y}_i)^2$ | | | |
| **Error** | $\Sigma(y_i - \hat{y}_i)^2$ | | | |
| **Total** | $\Sigma(y_i - \bar{y}_i)^2$ | | | |

# Three kinds of variance

- With these three types of error, does it remind us of anything?
- How about an ANOVA table?

| Source | Sums of Squares | df | Mean Squares | F-value |
|---|---|---|---|---|
| **Regression** | $\Sigma(\hat{y}_i - \bar{y}_i)^2$ | ? | | |
| **Error** | $\Sigma(y_i - \hat{y}_i)^2$ | ? | | |
| **Total** | $\Sigma(y_i - \bar{y}_i)^2$ | ? | | |



# Three kinds of variance

- With these three types of error, does it remind us of anything?
- How about an ANOVA table?

| Source | Sums of Squares | df | Mean Squares | F-value |
|---|---|---|---|---|
| **Regression** | 32.98 | k | | |
| **Error** | 300.17 | n-k-1 | | |
| **Total** | 334.28 | n-1 | | |

# Three kinds of variance

## Actually filling it in

| Source | Sums of Squares | df | Mean Squares | F-value |
|---|---|---|---|---|
| **Regression** | 32.98 | 1 | ? | |
| **Error** | 300.17 | 217 | ? | |
| **Total** | 334.28 | 218 | | |



# Three kinds of variance

## Actually filling it in

| Source | Sums of Squares | df | Mean Squares | F-value |
|---|---|---|---|---|
| **Regression** | 32.98 | 1 | 32.98 | |
| **Error** | 300.17 | 217 | 1.38 | |
| **Total** | 334.28 | 218 | | |

# Three kinds of variance

## Actually filling it in

| Source | Sums of Squares | df | Mean Squares | F-value |
|---|---|---|---|---|
| **Regression** | 32.98 | 1 | 32.98 | 23.84 |
| **Error** | 300.17 | 217 | 1.38 | |
| **Total** | 334.28 | 218 | | |



# What Does RStudio tell us?

```
lm(Life.satisfaction ~ Social.support, unemp)
```

```
##
## Call:
## lm(formula = Life.satisfaction ~ Social.support, data = unemp)
##
## Coefficients:
##    (Intercept)  Social.support
##         2.2136          0.3855
```

# What Does RStudio tell us?

- Using R to generate an ANOVA

```
summary(aov(Life.satisfaction ~ Social.support, unemp))
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)    
## Social.support   1  32.98   32.98   23.84 2.03e-06 ***
## Residuals      217 300.17    1.38                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```



# What Does RStudio tell us?

- And the "more important" "Simple" Linear Model

```
ls_mod <- lm(Life.satisfaction ~ Social.support, unemp)
summary(ls_mod) #This summary() is the BACKBONE of our regression output
```

```
## 
## Call:
## lm(formula = Life.satisfaction ~ Social.support, data = unemp)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.5124 -0.5633  0.1459  0.7387  2.8750 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     2.21363    0.46822   4.728 4.08e-06 ***
## Social.support  0.38554    0.07896   4.883 2.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.176 on 217 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.09899,    Adjusted R-squared:  0.09484 
## F-statistic: 23.84 on 1 and 217 DF,  p-value: 2.028e-06
```

# What does it all mean?

- Our errors from the model (and from no-model), tell give use sums of squares
- These can be used to generate an F-statistic (as in an ANOVA table)
- This F-statistic tells us "how well the overall model fits the desired outcome variable."
- F is an "overall" model statistic



# Back to residals…

- Recall:
  - Our residuals are assumed to be independent
  - And normally distributed
  - And have a mean of zero
  - *And* have a variance of $\sigma_e^2$

$$\sigma_e^2 = \frac{\Sigma(y - \hat{y})^2}{(n - p)}$$

## Back to residals…

- Recall:
  - Our residuals are assumed to be independent
  - And normally distributed
  - And have a mean of zero
  - *And* have a variance of $\sigma_e^2$

$$\text{MSE} = \sigma_e^2 = \frac{\Sigma(y - \hat{y})^2}{(n - p)}$$



## Back to residals…

- Recall:
  - Our residuals are assumed to be independent
  - And normally distributed
  - And have a mean of zero
  - *And* have a variance of $\sigma_e^2$

$$\text{RMSE} = \sqrt{\sigma_e^2} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{(n - p)}}$$

# RMSE...

- "Calculate" by hand:

```
sse <- sum(unemp$res_sq, na.rm = TRUE)
mse <- (sse / 437)
mse
```

```
## [1] 0.6868841
```

```
sqrt(mse)
```

```
## [1] 0.8287847
```



# RMSE...

- "Calculate" by hand:

```
sse <- sum(unemp$res_sq, na.rm = TRUE)
mse <- (sse / 97)
mse
```

```
## [1] 3.094519
```

```
sqrt(mse)
```

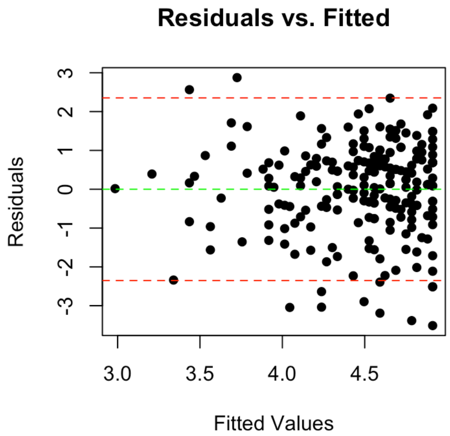
```
## [1] 1.759124
```

- Ask R to help:

```
summary(ls_mod)$sigma
```

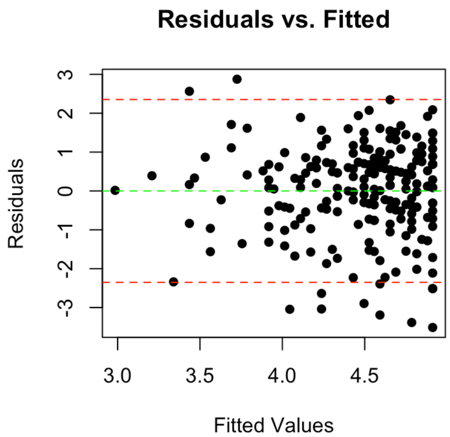
```
## [1] 1.176123
```

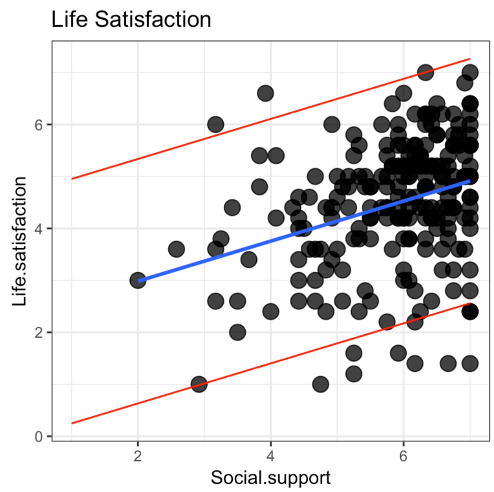
# MSE Visually:






# MSE Visually:



# Remember:

```
ls_mod <- lm(Life.satisfaction ~ Social.support, unemp)
ls_mod
```

```
##
## Call:
## lm(formula = Life.satisfaction ~ Social.support, data = unemp)
##
## Coefficients:
##    (Intercept)  Social.support
##         2.2136          0.3855
```



# And this is what happens

Our Population Linear Model:

$$\hat{y} = \beta_0 + \beta_1 x$$

Our *Inferred* Linear Model:

$$\hat{y} = b_0 + b_1 x$$

- We can assume that $b_0$ is an unbiased estimator of $\beta_0$
- We can assume that $b_1$ is an unbiased estimator of $\beta_1$

## And this is what happens

- And we can get variances for both $b_0$ and $b_1$:

$$Var(b_0) = \sigma_e^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2} \right]$$

$$Var(b_1) = \frac{\sigma_e^2}{\Sigma(x_i - \bar{x})^2}$$



## Now to Standard Error

- How do we *classically* go from Variance to Standard Error (or Standard Deviation)?

# Now to Standard Error

- How do we *classically* go from Variance to Standard Error (or Standard Deviation)?

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

$$S = \sqrt{S^2}$$



# Now to Standard Error

## Intercept

$$Var(b_0) = \sigma_e^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2} \right]$$

$$s.e.(b_0) = \hat{\sigma}_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}}$$

# Now to Standard Error

Slope

$$Var(b_1) = \frac{\sigma_e^2}{\Sigma(x_i - \bar{x})^2}$$

$$s.e.(b_1) = \frac{\hat{\sigma}_e}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$



# And now we have our grail

- What test did we use for the significance of Pearson Correlation?

# And now we have our grail

- With standard error in hand, we can now perform our test of inference for *both* the intercept and the slope.
- For the slope:

$$H_0 : b_1 = 0 \; H_1 : b_1 \neq 0$$

$$t_1 = \frac{b_1}{s.e.(b_1)}$$

Now, evaluated with a t-distribution with $(n - p)$ df (the number of observations minus the number of regression coefficients). We can also use $(n - k - 1)$ df.



# And the intercept too!

- For the intercept:

$$H_0 : b_0 = 0$$
$$H_1 : b_0 \neq 0$$

$$t_0 = \frac{b_0}{s.e.(b_0)}$$

Again, evaluated with a t-distribution with $(n - p)$ df (the number of observations minus the number of regression coefficients). We can also use $(n - k - 1)$ df.

# In action:

```
summary(ls_mod)
```

```
##
## Call:
## lm(formula = Life.satisfaction ~ Social.support, data = unemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5124 -0.5633  0.1459  0.7387  2.8750
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.21363    0.46822   4.728 4.08e-06 ***
## Social.support  0.38554    0.07896   4.883 2.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.176 on 217 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.09899,    Adjusted R-squared:  0.09484
## F-statistic: 23.84 on 1 and 217 DF,  p-value: 2.028e-06
```



# What does it all mean?

- Our estimate of sigma gets us estimates for the s.e. for both $b_0$ and $b_1$
- This gets us a way to test both simple regression parameters
- This will come in handy as we add additional independents
- F-statistic: Overall model
- t-statistic: Individual parameter = to 0
  - In the case of simple regression: does our simple slope differ from zero

# And Pearson too!

- Remember, we can also evaluate the value of r with a t-distribution as well:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

or:

$$t = \frac{r-0}{\sqrt{(1-r^2)/(n-2)}}$$

```
library(psych)
vars <- c("Life.satisfaction", "Social.support")
corr.test(select(unemp, one_of(vars)), use="pairwise.complete.obs")$t
```

```
##                     Life.satisfaction Social.support
## Life.satisfaction                 Inf       4.882719
## Social.support               4.882719            Inf
```



# What does it all mean?

- Our estimate of sigma gets us estimates for the s.e. for both $b_0$ and $b_1$
- This gets us a way to test both simple regression parameters
- This will come in handy as we add additional independents
- F-statistic: Overall model
- t-statistic: Individual parameter = to 0
  - In the case of simple regression: does our simple slope differ from zero
- t-statistic: A Pearson Correlation compared to zero.