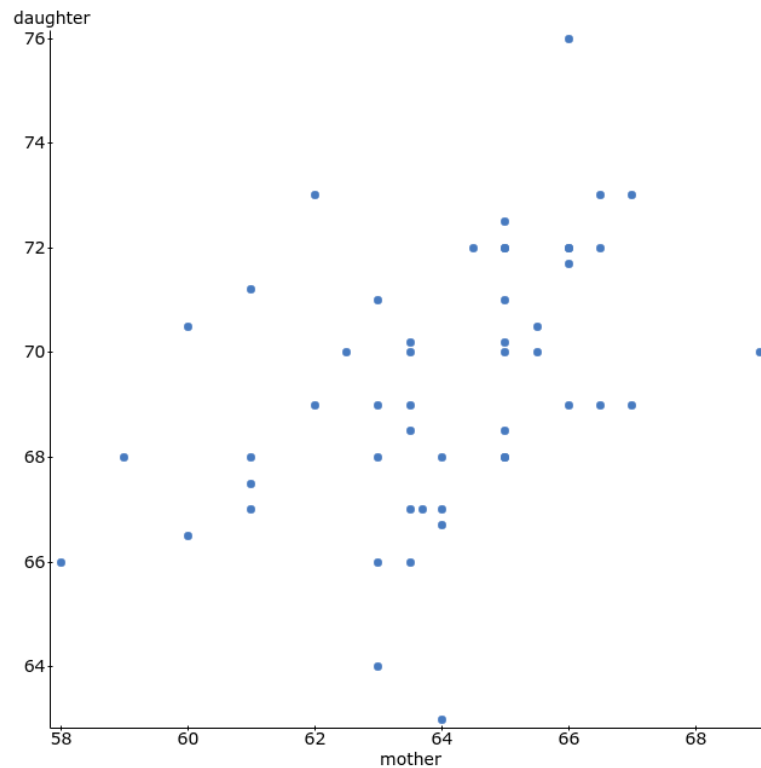# Group Work - Week 11

1 The file "Galton-mother-daughter.csv" contains a subset 50 subjects from Galton's mother/daughter height data.

(a) Create a scatterplot of the data. Does there appear to be a linear relationship between mother's heights and daughter's heights?

There does appear to be a linear relationship.

(b) Conduct a correlation hypothesis test at $\alpha = 0.05$ significance level. If there is significant correlation, how would you describe the strength of the correlation?

From Summary Stats $\rightarrow$ Correlation:

$r = 0.427$, $p = 0.002 < \alpha = 0.05$. **Reject H_0.**
There is evidence that heights of mothers and daughters are correlated.
Mother and daughter heights are moderately correlated.

(c) Find the estimated regression line for the relationship between mother's heights (predictor variable) and daughter's heights (response variable)? Is the slope significantly different than zero?

From Regression → Simple Linear:

Simple linear regression results:
Dependent Variable: daughter
Independent Variable: mother
daughter = 38.235282 + 0.48705835 mother
Sample size: 50
R (correlation coefficient) = 0.42728859
R-sq = 0.18257554
Estimate of error standard deviation: 2.3094977

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	38.235282	9.5144322	≠ 0	48	4.0186615	0.0002
Slope	0.48705835	0.14875203	≠ 0	48	3.2742973	0.002

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	57.183573	57.183573	10.721023	0.002
Error	48	256.02143	5.3337797		
Total	49	313.205			

$\hat{y} = 38.24 + 0.49x$

$t = 3.274$, $p = 0.002$. **The slope is significantly different than zero.**

(d) What is the best predicted daughter's height for a mother that is 56 inches tall? Is it appropriate to make such a prediction?

Since we have a significant correlation, use the regression equation for the prediction.
$\hat{y} = 38.24 + 0.49(56) = 65.68$
Or from StatCrunch, $\hat{y} = 65.51$

Mothers' heights range from 58 to 69 inches. Thus, making a prediction for a mother's height of 56 inches is not appropriate.

3   The file "MCA_scores_17.csv" on D2L contains average math MCA scores for 11th graders in 2017 by MN public school district, as well as percentage of 11th graders receiving free lunches in the district. Districts with missing data and charter schools are excluded.

(a) Create a scatterplot of the data. Does there appear to be a linear relationship between percentage of students receiving free lunch and average MCA scores?

There does appear to be a linear relationship.
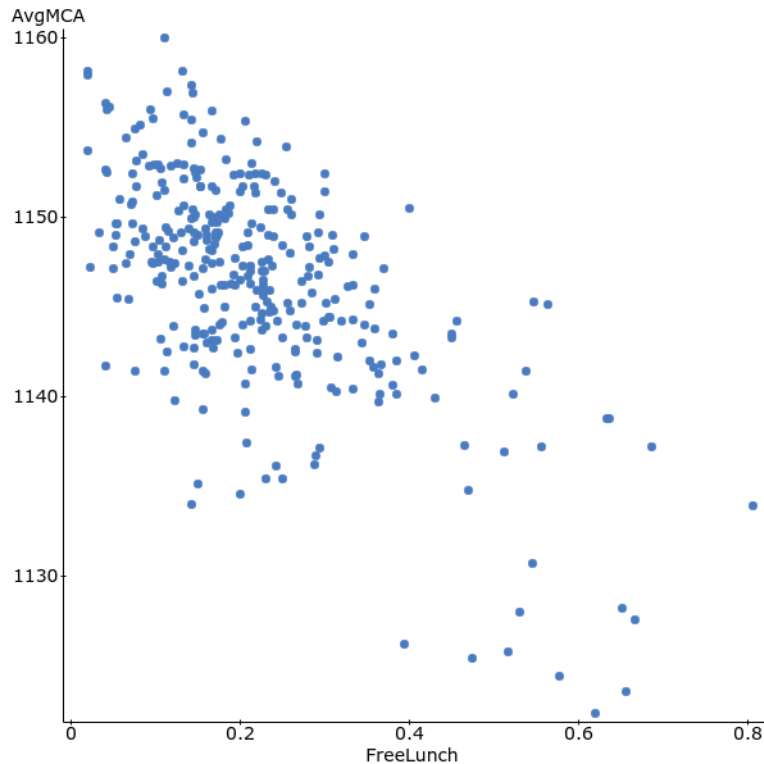
(b) Conduct a correlation hypothesis test at $\alpha = 0.01$ significance level. If there is significant correlation, how would you describe the strength of the correlation?

From Summary Stats → Correlation:

$r = -0.652$, $p < 0.0001 < \alpha = 0.01$. **Reject H_0.**
There is evidence that proportion of free lunch students in a district and average MCA scores are correlated.
Proportion of free lunch students in a district and average MCA scores are moderately correlated, close to highly correlated.

(c) Find the estimated regression line for the relationship between percentage of students receiving free lunch (predictor variable) and average MCA scores (response variable)? Is the slope significantly different than zero?

From Regression → Simple Linear:

Simple linear regression results:
Dependent Variable: AvgMCA
Independent Variable: FreeLunch
AvgMCA = 1153.0699 - 30.172593 FreeLunch
Sample size: 323
R (correlation coefficient) = -0.65192818
R-sq = 0.42501035
Estimate of error standard deviation: 4.6164296

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	1153.0699	0.50424373	≠ 0	321	2286.7313	<0.0001
Slope	-30.172593	1.9588013	≠ 0	321	-15.403601	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	5056.5807	5056.5807	237.27092	<0.0001
Error	321	6840.9666	21.311422		
Total	322	11897.547			

$\hat{y} = 1153.07 - 30.17x$

$t = -15.4$, $p < 0.0001$. **The slope is significantly different than zero.**

(d) What is the best predicted average MCA score for a district that has 45% of 11th grade students receiving free lunch? Is it appropriate to make such a prediction?

Since we have a significant correlation, use the regression equation for the prediction.
$\hat{y} = 1153.07 - 30.17(0.45) = 1139.49$
Or from StatCrunch, $\hat{y} = 1139.49$

Proportions of free lunch students range from 0.0202 to 0.8046. Thus, making a prediction for a free lunch proportion of 0.45 is appropriate.