Relationship between two numerical variables	
Correlation	Definition: Linear association between two variables, denoted as R Properties 1. Magnitude of R measures strength of linear association 2. Sign of R indicate slope 3. R is between -1 and 1 4. R is unitless (not affect by changes in center or scale such as unit conversion) 5. Correlation of X with Y = Y with X 6. R is sensitive to outliers
Residuals	Definintion: The sum of the difference between the data points and the linear regression. ▸ difference between the observed and predicted y **residual:** $e_i = y_i - \hat{y}_i$ Interpret: - Positive residual: observed value higher than predicted - Negative residual: observed value is lower than predicted
Least Squares Line (line of best fit)	To minimize residuals. Calculated by finding the slope, x-intercept, y-intercept. Features: ● Slope: For each unit increase in x, the y is expected to be higher/lower on average by the slope ● X-intercept - serves to adjust the height of the line. Slope: **slope:** $b_1 = \frac{s_y}{s_x} R$ s_x : SD of x, s_y : SD of y, $R = cor(x, y)$ Notation:

	parameter	point estimate
intercept	β_0	b_0
slope	β_1	b_1

Least Squares Line Example:

The standard deviation of % living poverty is 3.1%, and the standard deviation of % HS graduates is 3.73%. Given that the correlation between these variable is -0.75, what is the slope of the regression line for predicting % living poverty from % HS graduates?

Step 1: Identify what's given

$s_y = 3.1\%$

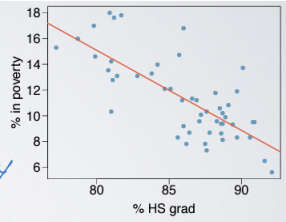
$s_x = 3.73\%$

$R = -0.75$

Step 2: Find the Slope

$b_1 = \frac{s_y}{s_x} R = \frac{3.1}{3.73} \times -0.75 \approx -0.62$

For each % point increase in H/S graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.



Step 3: Estimate the regression using points given

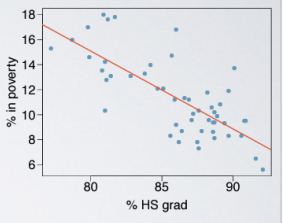
Given that the average % living in poverty is 11.35%, and the average % HS graduates is 86.01%, what is the intercept of the regression line for predicting % living poverty from % HS graduates?

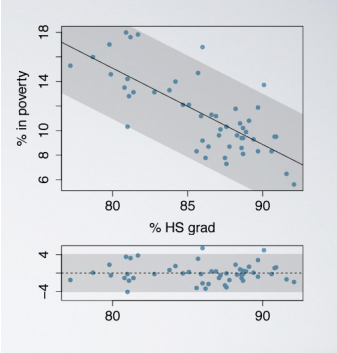
$\bar{y} = 11.35\%$

$\bar{x} = 86.01\%$

$b_0 = \bar{y} - b_1 \bar{x} = 11.35 - (-0.62)\ 86.01 = 64.68$

States with no H/S graduates are expected on average to have 64.68% of their residents living below the poverty line.



Purpose of Least Squares Line	Prediction - Predict the value of the response variable for a given value of the explanatory variable Extrapolation - Applying a model estimate to values outside of the realm of the original data - Sometimes the intercept might be an extrapolation
Conditions for Linear Regression	https://gallery.shinyapps.io/slr_diag/ Linearity - Linear relationship - Methods for fitting a model to non-linear relationships exist - Check using a scatter plot of the data or a residual plot. Nearly normal residuals - Normally distributed centered at 0 - May not be satisfied if there are unusual observations (outliers) - Check using a histogram or normal probability plot Constant Variability - Variability of points around regression line should be roughly constant - Implies that the variability of residuals around 0 line should be roughly constant as well - Aka homoscedasticity - Check using a residuals plot  -
R squared	Definition: - Strength of the fit of a linear model calculated as the square of the correlation coefficient - Tells us what percent of variability in the response variable is explained by the model - The rest of the variability is explained by variables not included in the model - Always between 0 and 1 Example: R^2 =0.5625 - Means 56.25% of the variability in the % of residents living in poverty among the states is explained by the model

Regression with Categorical variables	Set 0 and 1 for each categorical variable, if it's the one you want, set the rest to 0 and the desired variable to 1.

Next, we use a new region variable (**region4**) with four levels: northeast, midwest, west, south. Write the linear regression model based on the regression output below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.00

% in poverty = 9.50 + 0.03 reg4:mw + 1.79 reg4:w + 4.16 reg4:s

Calculate the predicted poverty rate for western states.

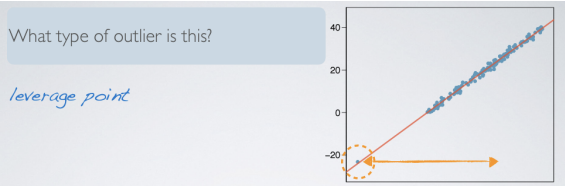
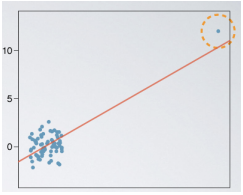
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.00

% in poverty = 9.50 + 0.03 reg4:mw + 1.79 reg4:w + 4.16 reg4:s

(reg4:mw = 0, reg4:w = 1, reg4:s = 0)

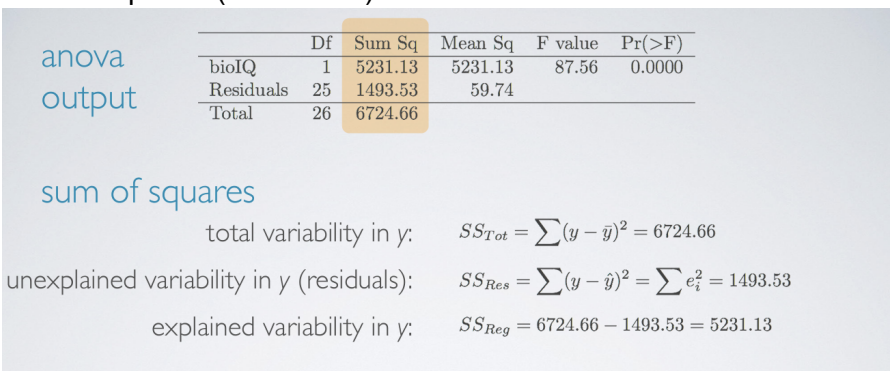
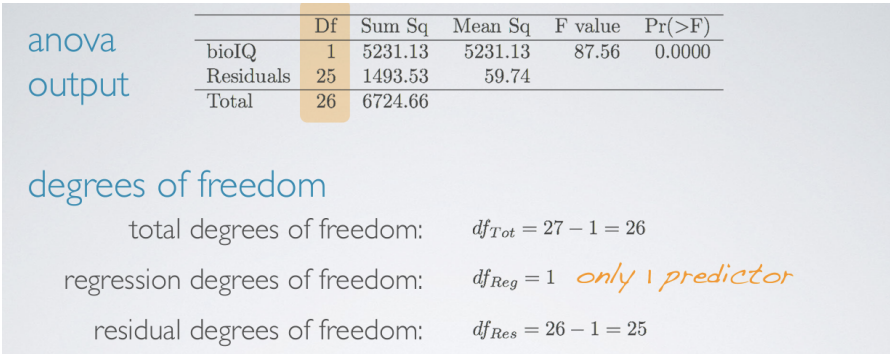
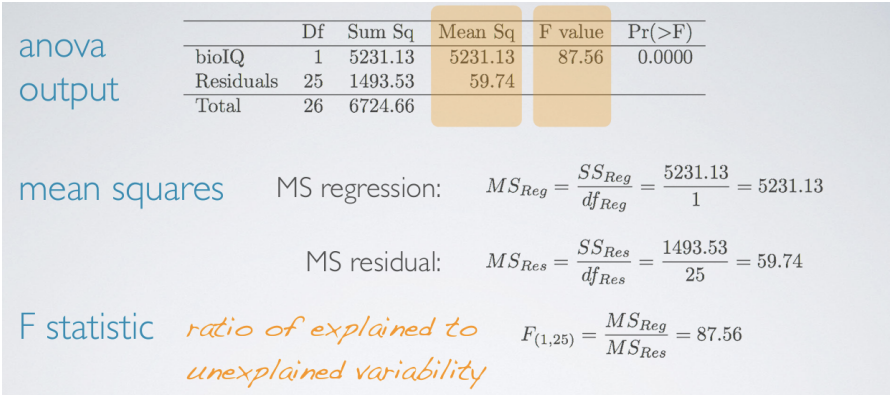
= 9.50 + 0 + 1.79 + 0

= 11.29

Outliers and Regression	
Outliers in Regression	Two Types: 1. Leverage Points: outliers that fall horizontally away from the center of the cloud but doesn't influence the slop of the regression line  2. Influential Points: influence the slope of the regression line a. Usually high leverage points 

Hypothesis Testing for Linear Regression	Hypothesis - H0 = slope is 0 (no relationship) - HA = slope != 0, there is a relationship T-statistic and df for Slope Calculation - t-statistic for the slope: $T = \frac{b_1 - 0}{SE_{b_1}}$ $df = n - 2$ - Df = n-2 because we are estimating for two parameters, b1 and b2 Confidence Interval for Slope - $b_1 \pm t^*_{df} SE_{b_1}$ - SE usually given Regression output gives b1, SEb1, two-tailed p-value for the t-test for the slope where the null value is 0
Hypothesis Test for Linear Regression Example:	Calculate the 95% confidence interval for the slope of the relationship between biological and foster twins' IQs? |	Estimate	Std. Error	t value	Pr(>|t|)
---	---	---	---	---
(Intercept)	9.2076	9.2999	0.99	0.3316

bioIQ	0.9014	0.0963	9.36	0.0000	
					
					 
					(Also given that 27 subjects were studied)
					
					Step 1: Find df
					- $df = 27 - 2 = 25$
					Step 2: Find t*
					- By hand…using t-table locate 0.05 for two tail and degree of freedom of 25
					- By R…
					- R
					`> qt(0.025, df = 25)`
					`[1] -2.059539`
					Step 3: Find the Confidence Interval
					- $0.9014 \pm 2.06 \times 0.0963 = (0.7, 1.1)$
					Step 4: Interpret
					- We are 95% confident that for each additional point on the biological twins' IQs, the foster twins' IQs are expected on average to be higher by 0.7 to 1.1 points.

Variability Partitioning (ANOVA)	Partitioning variability is the process of breaking down the total variability in a dataset into different components that explain the observed data. Use ANOVA in context of linear regression Sum of Squares (Calculated)  Degrees of Freedom  Mean Squares and F-Statistic  P-Value

anova

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

small p-value → reject H_0

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

The data provide convincing evidence that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

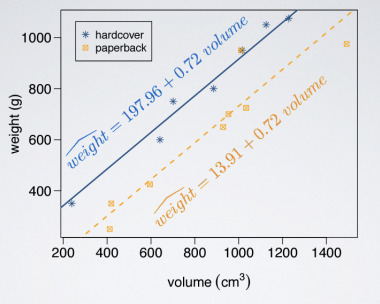
R^2 also derived from Proportion of Sum of Squares (remember, the other way is the square of correlation coefficient)

$$(2) \quad R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{SS_{Reg}}{SS_{Tot}} = \frac{5231.13}{6724.66} \approx 0.78$$

Multiple Regressions

Multiple Predictors

We are comparing hardcover and paperback books



Plug into R to find the linear model output (lm()). ANOVA is referencing hardcover as the base and then layers on the paperback to do the paperback calculation.

The '+cover' adds on the second item to be considered

```
# fit model
> book_mlr = lm(weight ~ volume + cover, data = allbacks)
> summary(book_mlr)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

This means that to look at ONLY hardcover, must plug in 0 for the cover:pb

	▸ For hardcover books: plug in **0** for cover: $\widehat{weight} = 197.96 + 0.72\,volume - 184.05 \times 0$ $= 197.96 + 0.72\,volume$ To look at ONLY paperback, must plug in 1 for the cover:pb to layer on the difference ▸ For paperback books: plug in **1** for cover: $\widehat{weight} = 197.96 + 0.72\,volume - 184.05 \times 1$ $= 13.91 + 0.72\,volume$ We can predict the weight of a paperback book that is 600cm^3 in volume. $197.96 + 0.72 \times 600 - 184.05 \times 1 = 445.91$ grams **Interpret the output:** Slope - Slope of Volume: for each 1cm^3 increase in volume, the model predicts the books to be heavier on average by 0.72grams - Slope of Cover:pb: paperback books weigh 184.05g lower than hardcover books Intercept - Hardcover books with n volume are expected to weigh 198g -> makes no sense, just used to adjust the height of the line Note that the model assumes the same slope for both hardcover and paperback. If that isn't' reasonable, we would include an interaction variable in the model.
Adjusted R-squared	Definition: Applies a penalty to R-squared for the number of predictors included in the model. The magnitude of the penalty depends on how k compares to our sample size n. The larger the sample size, the more predictors the model can handle and the less penalty is going to be for additional predictors being added to the mode. **adjusted R^2:** $R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1}\right)$ k : number of predictors Example:

predicting poverty from % female householder + % white

R

```
> pov_mlr = lm(poverty ~ female_house + white, data = states)
> summary(pov_mlr)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

R

```
> anova(pov_mlr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2 = \frac{132.57 + 8.21}{480.25} = 0.29$$

We've partitioned the predicted poverty level into two predictors: female householder and white, decreasing the residual (unexplained). However, we need to adjust the R^2 for the additional predictor.

Calculate adjusted R^2 for the multiple linear regression model predicting % living in poverty from % female householders and % white. Remember n = 51 (50 states + DC).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1}\right)$$

$$= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1}\right) = 0.26$$

As a result, adjusted R^2 is less than the original R^2.

Why is this important?

- When any variable is added to the model, R^2 increases but that's not necessary true.
- If the added variable doesn't really provide any new information or is completely unrelated, the adjusted R^2 does not increase
- Therefore we choose adjusted R^2 for the penalty it includes. We choose models with the highest adjusted R^2.

Collinearity and Parsimony

Collinearity

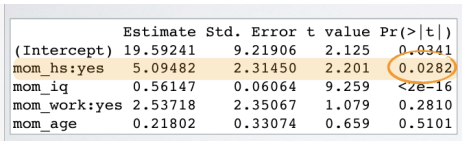
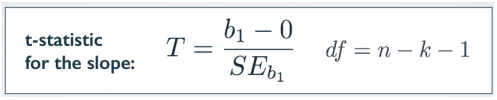
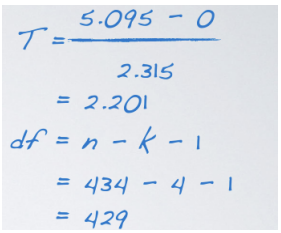
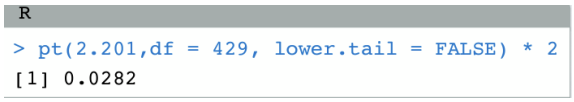
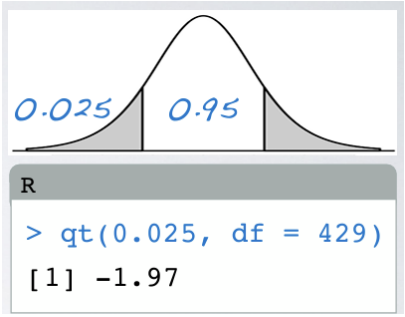
- Two predictors are correlated with each other, which complicates a model estimation because predictors should be independent
- The addition of collinear variables can result in biased estimates of the regression parameters.

Parsimony

	- Want to avoid adding predictors associated with each other because it would bring nothing new to the table so we prefer the simplest best model: ***parsimonious model*** - Occam's razor: among competing hypotheses, the one with the fewest assumption should be selected

Inference for Multiple Regression and Model Selection	

Inference for Multiple Linear Regression (MLR)	Using R: We have a set of data and we want to know if different factors (explanatory variables) for the moms impact their kid's IQ score.
	
	Using R, we input the full model
	`# full model`
	`> cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive)`
	`> summary(cog_full)`
	
	This yields the output below:
	`Coefficients:`
	`Estimate Std. Error t value Pr(>|t|)`
	`(Intercept) 19.59241 9.21906 2.125 0.0341 *`
	`mom_hs:yes 5.09482 2.31450 2.201 0.0282 *`
	`mom_iq 0.56147 0.06064 9.259 <2e-16 ***`
	`mom_work:yes 2.53718 2.35067 1.079 0.2810`
	`mom_age 0.21802 0.33074 0.659 0.5101`
	
	`Residual standard error: 18.14 on 429 degrees of freedom`
	`Multiple R-squared: 0.2171, Adjusted R-squared: 0.2098`
	`F-statistic: 29.74 on 4 and 429 DF, p-value: < 2.2e-16`
	
	Step 1: Take a F-test to determine the model as a whole to know if it is worthwhile to hypothesis test for slopes
	
	- Set the Hypothesis for F-test
	  - $H_0: \beta_1 = \beta_1 = \ldots = \beta_k = 0$
	H_A: At least one β_i is different than 0
	`F-statistic: 29.74 on 4 and 429 DF, p-value: < 2.2e-16`
	  - H0=all the slopes are 0, there is no relation
	  - HA:=atleast one slope is not 0
	- Analysis:
	  - Since the p-value is samaller, we reject the null hypothesis and know that at least one slope is not 0
	Step 2: Perform hypothesis testing (T-Test) on each of the factors to determine whether or not they are a significant predictor of kids IQ. Example below is with mom who went to high school.
	
	- Set the Hypothesis for T-Test
	  - $H_0: \beta_1 = 0$, when all other variables are included in the model
	$H_A: \beta_1 \neq 0$, when all other variables are included in the model

	- H0= slope is 0, no relation - HA= slope is not 0, there is a relationship - Look at the p-value to see if we can reject the null hypothesis or not -  - Mom_hs, mom_iq are significant - mom_work , mom_age are not significant
Inference for Multiple Linear Regression (MLR) Done by Hand	Verify the T score and the p-value for the slope of **mom_hs**. Step 1: T-score and df - Formula -  -  - Output -  Calculate the 95% confidence interval for the slope of **mom_work**. Step 1: Find critical point t* and df - Using R/t-table -  Step 2: Find the confidence interval

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hs:yes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_work:yes	2.53718	2.35067	1.079	0.2810
mom_age	0.21802	0.33074	0.659	0.5101

t-statistic for the slope: $T = \frac{b_1 - 0}{SE_{b_1}}$ $df = n - k - 1$

$$T = \frac{5.095 - 0}{2.315} = 2.201$$

$$df = n - k - 1 = 434 - 4 - 1 = 429$$

R

```
> pt(2.201,df = 429, lower.tail = FALSE) * 2
[1] 0.0282
```


R

```
> qt(0.025, df = 429)
[1] -1.97
```

	- 		Estimate	Std. Error	t value	Pr(>|t|)
---	---	---	---	---
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hs:yes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_work:yes	2.53718	2.35067	1.079	0.2810
mom_age	0.21802	0.33074	0.659	0.5101	 Residual standard error: 18.14 on 429 degrees of freedom - $2.54 \pm 1.97 \times 2.35 \approx (-2.09, 7.17)$ Step 3: Interpret - We are 95% confident tat, all else being equal, the model predicts that children whose moms worked during the first 3 years of their lives score 2.09 points lower to 7.17 points higher than those whose mom did not work.
Model Selection - Stepwise Model Selection	**Backwards Elimination - adjusted R^2** - Start with the full model - Drop one variable at a time and record adjusted r^2 of each smaller model - Pick the model with the highest increase in adjusted R^2 - Repeat until none of the models yield an increase in adjusted R^2 	step	variables included	removed	adjusted R
---	---	---	---
FULL	kid_score ~ mom_hs + mom_iq + mom_work + mom_age		0.2098
STEP 1	kid_score ~ mom_iq + mom_work + mom_age	[-mom_hs]	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	[-mom_iq]	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	[-mom_work]	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	[-mom_age]	0.2109
STEP 2	kid_score ~ mom_iq + mom_work	[-mom_hs]	0.2024
	kid_score ~ mom_hs + mom_work	[-mom_iq]	0.0546
	kid_score ~ mom_hs + mom_iq	[-mom_work]	0.2105 | **Backwards Elimination - p-value** - Start with the full model - Drop the variable with the highest p-value and refit a smaller model - Repeat until all variables left in the model are significant backwards elimination - p-value FULL |	Estimate	Std. Error	t value	Pr(>|t|)
---	---	---	---	---
(Intercept)	19.5924	9.2191	2.13	0.0341
mom_hs:yes	5.0948	2.3145	2.20	0.0282
mom_iq	0.5615	0.0606	9.26	0.0000
mom_work:yes	2.5372	2.3507	1.08	0.2810
~~mom_age~~	~~0.2180~~	~~0.3307~~	~~0.66~~	~~0.5101~~ | STEP 1 |	Estimate	Std. Error	t value	Pr(>|t|)
---	---	---	---	---
(Intercept)	24.1794	6.0432	4.00	0.0001
mom_hs:yes	5.3823	2.2716	2.37	0.0183
mom_iq	0.5628	0.0606	9.29	0.0000
~~mom_work:yes~~	~~2.5664~~	~~2.3487~~	~~1.09~~	~~0.2751~~ | STEP 2 |	Estimate	Std. Error	t value	Pr(>|t|)
---	---	---	---	---
(Intercept)	25.7315	5.8752	4.38	0.0000
mom_hsyes	5.9501	2.2118	2.69	0.0074
mom_iq	0.5639	0.0606	9.31	0.0000	 **Forward Selection - adjusted R^2**

	- Start with a single predictor regression of the response vs explanatory variable - Pick he model with the highest adjusted R^2 - Add the remaining variables one at a time to the existing model and pick the one with highest adjusted R^2 - Repeat until addition of variables does not result in higher adjusted R^2 **Forward Selection - p-value** - Start with a single predictor regression of the response vs explanatory variable - Pick the variable with lowest p-value - Add the remaining variable one at a time to the existing model and pick the variable with lowest p-value
Model Selection - Backward Elimination Example	**NOTE:** if you have categorical variables with various levels and you are doing backwards elimination. Do not drop variable Race because the level asian's p-value is significant
Model Selection - R^2 vs P-value	Adjusted R^2 is more reliable prediction but p-value has more significant predictors

step	variables included	adjusted R
STEP 1	kid_score ~ mom_hs	0.0539
	kid_score ~ mom_work	0.0097
	kid_score ~ mom_age	0.0062
	kid_score ~ mom_iq	0.1991
STEP 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_iq + mom_age	0.1999
	kid_score ~ mom_iq + mom_hs	0.2105
STEP 3	kid_score ~ mom_iq + mom_hs + mom_age	0.2095
	kid_score ~ mom_iq + mom_hs + mom_work	0.2109
STEP 4	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	0.2098

The following model uses data from the American Community Survey to predict income from hours worked per week, race, and gender. Which variable (if any) should be dropped from the model first when doing backwards elimination using the p-value approach?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2782.5726	6676.5534	0.42	0.6770
hrs_work	1247.2128	146.2013	8.53	0.0000 ✓
race:black	-9565.3090	6393.2168	-1.50	0.1350
race:asian	35816.6156	8690.3484	4.12	0.0000 ✓
race:other	-11112.8617	7213.3220	-1.54	0.1238
gender:female	-16430.0916	3803.4700	-4.32	0.0000 ✓

don't drop any variables

R

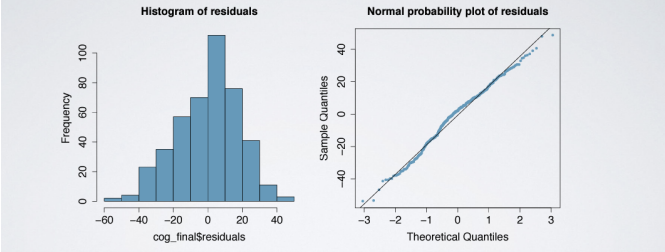
```
> cog_final = lm(kid_score ~ mom_hs + mom_iq + mom_work, data = cognitive)
> summary(cog_final)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.17944    6.04319   4.001 7.42e-05 ***
mom_hsyes    5.38225    2.27156   2.369   0.0183 *
mom_iq       0.56278    0.06057   9.291  < 2e-16 ***
mom_workyes  2.56640    2.34871   1.093   0.2751

Residual standard error: 18.13 on 430 degrees of freedom
Multiple R-squared:  0.2163, Adjusted R-squared:  0.2109
F-statistic: 39.57 on 3 and 430 DF,  p-value: < 2.2e-16
```

We selected this model using the adjusted R^2 method, which tells us that including the mom_workyes gives us higher predictive power even though it may not be statistically significant than the other variables (looking at p-values)

Conditions required for Multiple Linear Regressions to be Valid

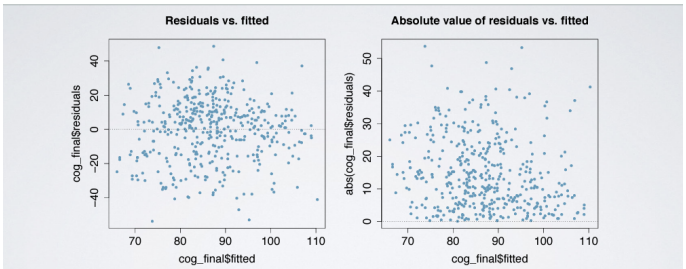
1. Linear relationship between numerical explanatory and response variable

```
> hist(cog_final$residuals)
> qqnorm(cog_final$residuals)
> qqline(cog_final$residuals)
```



2. Constant variability of residuals

```
> plot(cog_final$residuals ~ cog_final$fitted)
> plot(abs(cog_final$residuals) ~ cog_final$fitted)
```



3. Independent residuals (independent observations)

	R `> plot(cog_final$residuals)`