Lecture 19 - Regression: Inference, Outliers, and Intervals

Sta102 / BME102

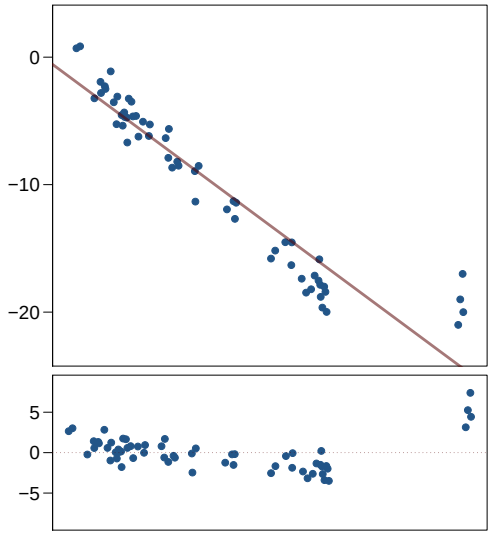
Colin Rundel

November 11, 2015

Types of outliers

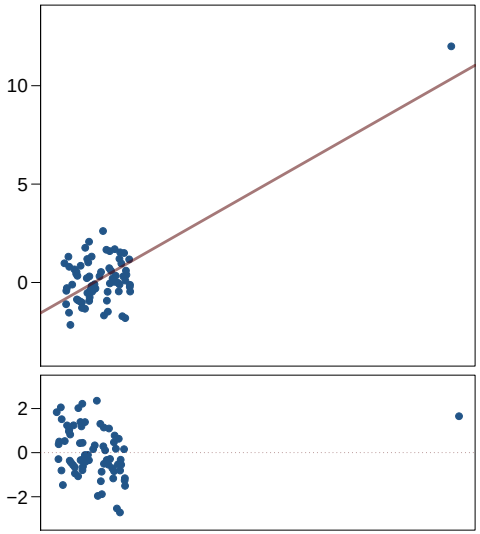
How does one or more outliers influence the least squares line?

To answer this question think of where the regression line would be with and without the outlier(s).



Types of outliers

How does the following outlier influence the least squares line?

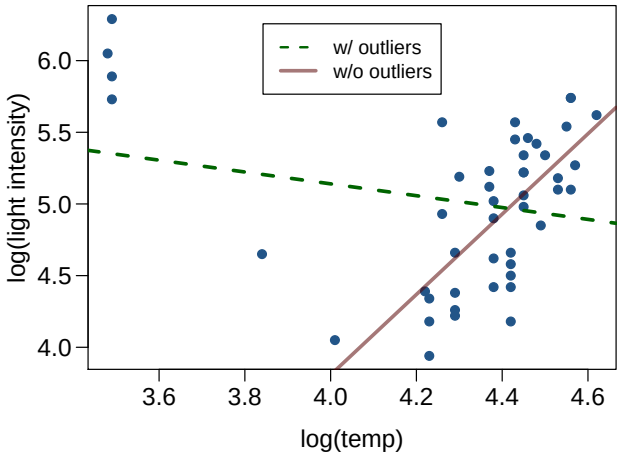


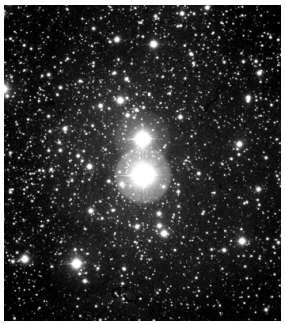
Some terminology

- *Outliers* are points that fall away from the cloud of points.
- Outliers that fall horizontally away from the center of the cloud are called *leverage* points.
- High leverage points that actually influence the slope of the regression line are called *influential* points.
- In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then it's not.

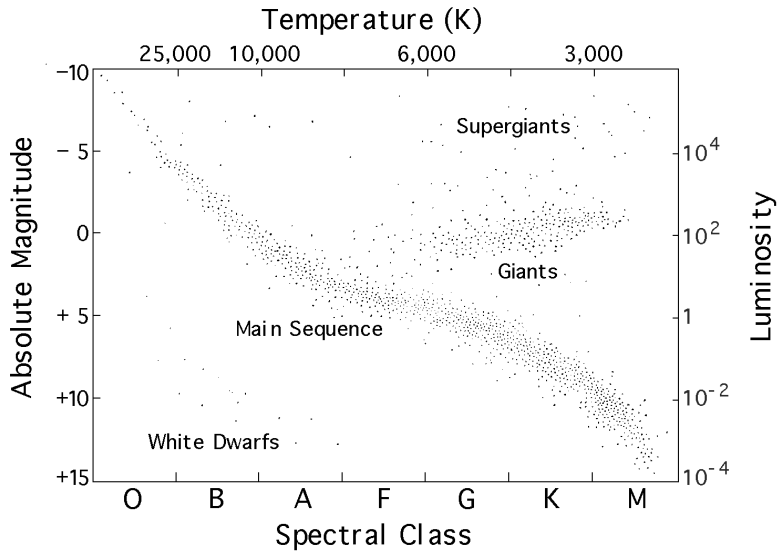
Influential points

Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.



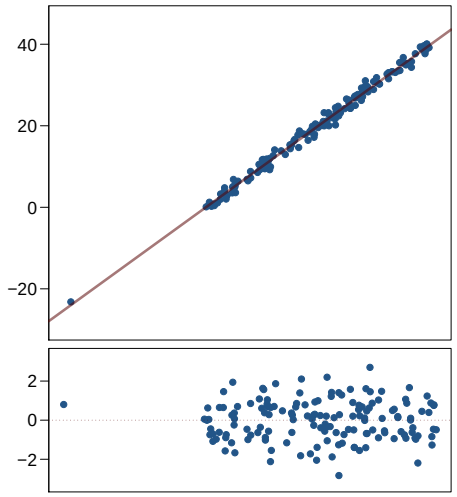


Hertzsprung-Russell Diagram



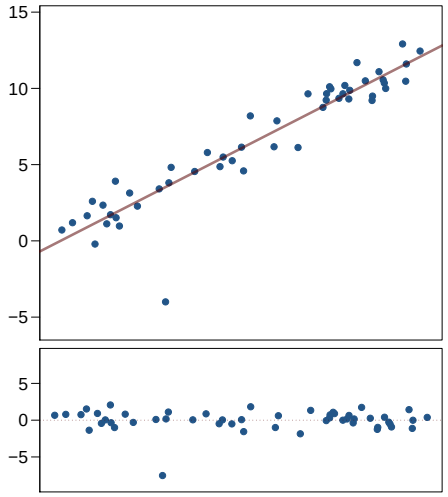
Types of outliers

Which type of outlier is displayed below?



Types of outliers

Which type of outlier is displayed below?



Recap

Are following statements true or false?

(1) Influential points always change the intercept of the regression line.

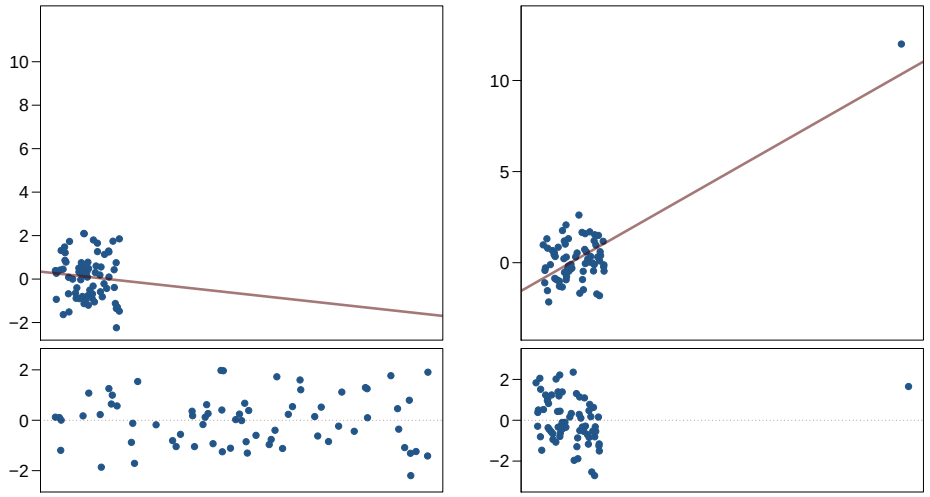
(2) Influential points always reduce R^2.

(3) It is much more likely for a high leverage point to be influential, than a low leverage point.

(4) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.

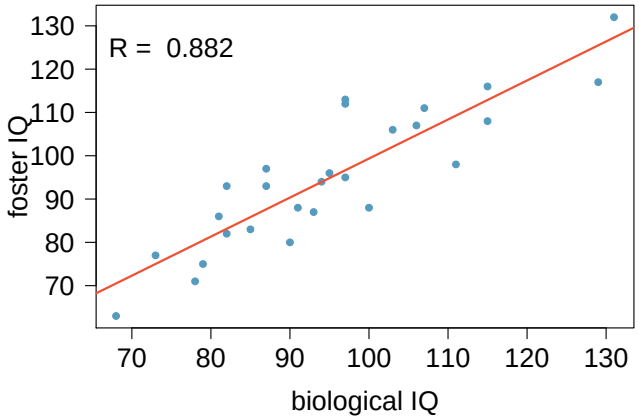
Recap (cont.)

$R = -0.091, R^2 = 0.0083$ $R = 0.72, R^2 = 0.522$



Nature vs. nurture?

In 1966 Cyril Burt published a paper called "The genetic determination of differences in intelligence: A study of monozygotic twins reared apart" The data consist of IQ scores for [an assumed random sample of] 27 identical twins, one raised by foster parents, the other by the biological parents.



Finding the regression line

	Foster IQ (y)	Biological IQ (x)
mean	$\bar{y} = 95.11$	$\bar{x} = 95.30$
sd	$s_y = 16.08$	$s_x = 15.73$
correlation	$R = 0.8819$	

$$b_1 = \frac{s_y}{s_x} R = \frac{16.08}{15.73} 0.8819 = 0.90$$

$$b_0 = \bar{y} - b_1 \bar{x} = 95.11 - 0.90\, 95.30 = 9.2$$

Regression Output

```
summary(lm(twins$Foster ~ twins$Biological))

## Call:
## lm(formula = twins$Foster ~ twins$Biological)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3512  -5.7311   0.0574   4.3244  16.3531
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.20760    9.29990   0.990    0.332
## twins$Biological  0.90144    0.09633   9.358  1.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.729 on 25 degrees of freedom
## Multiple R-squared: 0.7779, Adjusted R-squared: 0.769
## F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

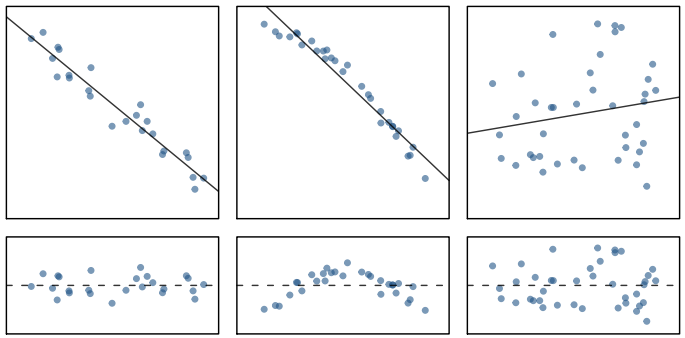

Conditions for inference

In order to perform inference, the following conditions must be met:

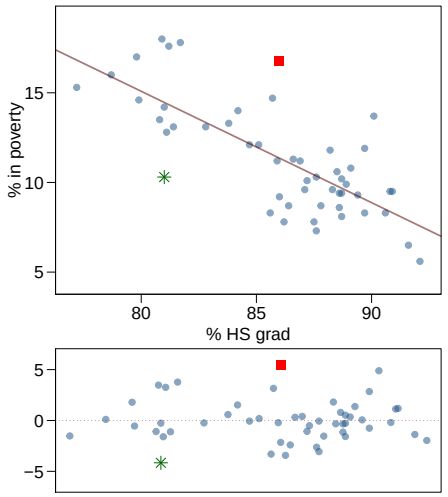
1. Linearity
2. Nearly normal residuals
3. Constant variability

Conditions: (1) Linearity

- The relationship between the explanatory and the response variable should be linear.
- Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class.
- Check using a *scatterplot* (x vs y) or a *residual plot* (x vs resid) .



Anatomy of a residuals plot



∗ *Rhode Island:*

$\%\ HS\ grad = 81 \qquad \%\ in\ poverty = 10.3$

$\widehat{\%\ in\ poverty} = 64.68 - 0.62 * 81 = 14.46$

$e_{RI} = \%\ in\ poverty - \widehat{\%\ in\ poverty}$

$= 10.3 - 14.46 = -4.16$

■ *Washington, DC:*

$\%\ HS\ grad = 86 \qquad \%\ in\ poverty = 16.8$

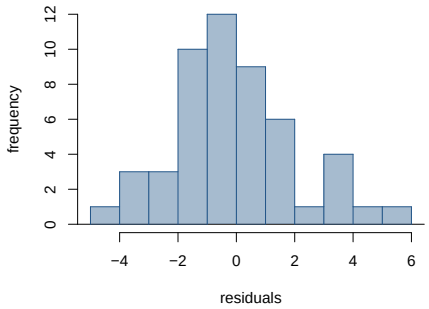
$\widehat{\%\ in\ poverty} = 64.68 - 0.62 * 86 = 11.36$

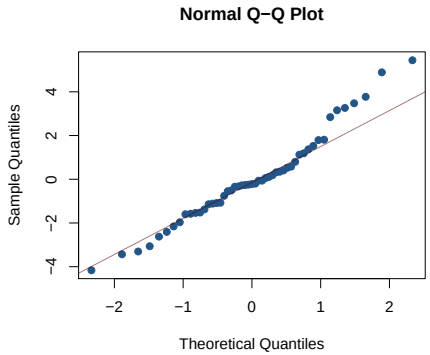
$e_{DC} = \%\ in\ poverty - \widehat{\%\ in\ poverty}$

$= 16.8 - 11.36 = 5.44$

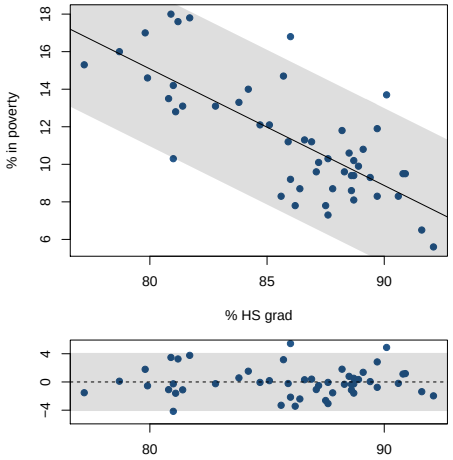
Conditions: (2) Nearly normal residuals

- The residuals should follow a nearly normal distribution.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Checked using a histogram or normal probability plot of residuals.





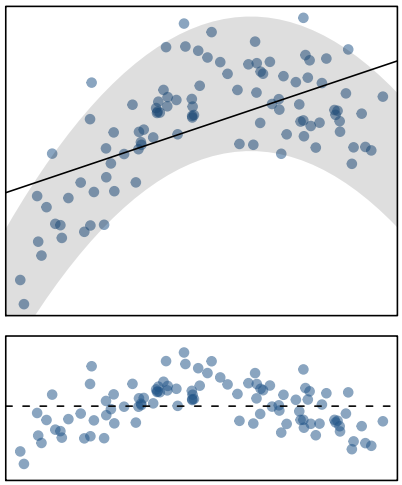
Conditions: (3) Constant variability



- The variability of the residuals from the least squares line should be constant.
- This implies that the variability of *any* region of the residual plot should be the same as any other region.
- Also called *homoscedasticity* / *heteroscedasticity*.
- Check using a residuals plot.

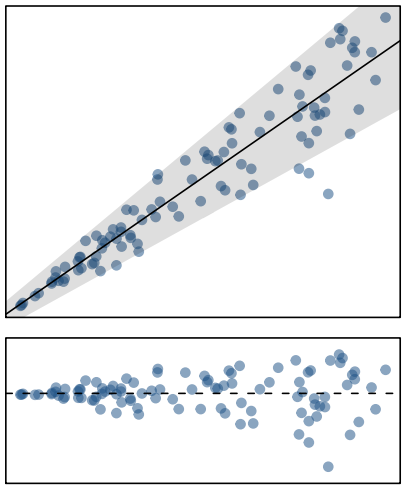
Checking conditions

What condition is this linear model violating?

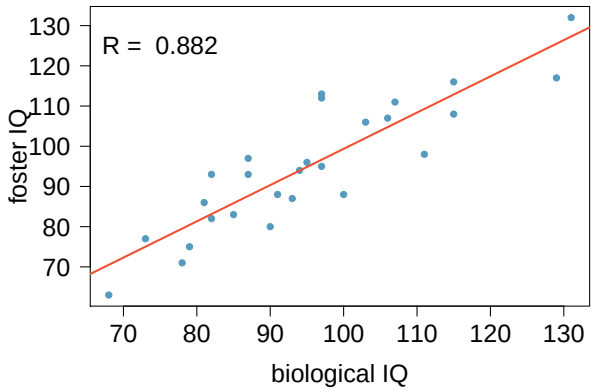


Checking conditions (II)

What condition is this linear model obviously violating?



Back to Nature vs nurture



	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

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin.

What are the appropriate hypotheses?

First consider what the null hypothesis should be, if there is no relationship between the two variables what value of the slope would we expect to see?

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Testing for the slope (cont.)

	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

- We always use a t-test in inference for regression parameters.
  Remember: *Test statistic,* $T = \frac{\text{point estimate} - \text{null value}}{SE}$
- Point estimate is b_1, the calculated slope for the observed data.
- SE_{b_1}, is the standard error associated with that slope.

$$SE_{b_1} = \frac{s_e}{\sqrt{n-1}\, s_x} = \frac{\sqrt{\sum_{i=1}^{n} \epsilon_i^2/(n-2)}}{\sqrt{n-1}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}}$$

- Degrees of freedom associated with the slope is $df = df_e = n - 2$, where n is the sample size.

Testing for the slope (cont.)

	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

$$\begin{aligned} T &= \frac{0.9014 - 0}{0.0963} = 9.36 \\ df &= 27 - 2 = 25 \\ \text{p-value} &= P(|t| > 9.36) < 0.01 \end{aligned}$$

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* $\pm$ *ME* and the degrees of freedom associated with the slope in a simple linear regression is $n - 2$. What is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	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

Recap

- Inference for the slope for a SLR model (only one explanatory variable):
  - Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2} \times SE_{b_1}$$

- The null value is usually 0, since we are usually checking for *any* relationship between the explanatory and the response variable.
- The regression output gives b_1, SE_{b_1}, and the *two-tailed* p-value for the t-test of the slope where the null value is 0.
- We rarely do inference on the intercept (since it is rarely meaningful), so we will be focus on the estimates and inference for the slope.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you have population data.
- If you have a sample that is non-random (biased), the results will be unreliable.
- The ultimate goal is to have independent observations – and you know how to check for those by now.

Variability partitioning

- We considered the t-test as a way to evaluate the strength of evidence for a hypothesis test for the slope of relationship between x and y.
- However, we can also consider the variability in y explained by x, compared to the unexplained variability.
- *Partitioning* the variability in y to explained and unexplained variability is something we have already done *(ANOVA)*.

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

The only change is that we use $\hat{y}_i$ instead of μ_i.

ANOVA output - Sum of squares

	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			

Sum of squares:

$$SS_{Tot} = \sum_i (y_i - \bar{y})^2 = 6724.66 \quad \textit{(total variability in y)}$$

$$SS_{Err} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

$$= 1493.53 \quad \textit{(unexplained variability in residuals)}$$

$$SS_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2$$

$$= SS_{Tot} - SS_{Err} \quad \textit{(explained variability in y)}$$

$$= 6724.66 - 1493.53 = 5231.13$$

Degrees of freedom:

$$df_{Tot} = n - 1 = 27 - 1 = 26$$

$$df_{Reg} = 2 - 1 = 1 \quad \textit{(there are 2 coefficients)}$$

$$df_{Res} = df_{Tot} - df_{Reg} = 26 - 1 = 25$$

ANOVA output - F-test

	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			

Mean sq.:

$$MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{5231.13}{1} = 5231.13$$

$$MS_{Err} = \frac{SS_{Err}}{df_{Err}} = \frac{1493.53}{25} = 59.74$$

F-statistic:

$$F_{(1,25)} = \frac{MS_{Reg}}{MS_{Err}} = 87.56 \quad \textit{(ratio of explained to unexplained variability)}$$

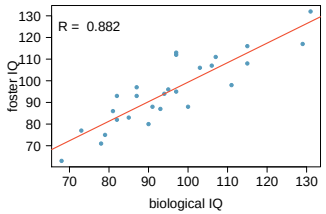
The null hypothesis is $\beta_0 = \beta_1 = 0$ and the alternative is $\beta_j \neq 0$ for some j. With a large F-statistic, and a small p-value, we reject H_0 and conclude that the linear model is significant.

Regression Output

```
summary(lm(twins$Foster ~ twins$Biological))

## Call:
## lm(formula = twins$Foster ~ twins$Biological)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3512  -5.7311   0.0574   4.3244  16.3531
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.20760    9.29990   0.990    0.332
## twins$Biological  0.90144    0.09633   9.358  1.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.729 on 25 degrees of freedom
## Multiple R-squared: 0.7779, Adjusted R-squared: 0.769
## F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```


ANOVA output - R^2 calculation



	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			

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{SS_{Reg}}{SS_{Tot}} = \frac{5231.13}{6724.66} = 0.7779$$

Sampling Distribution of $\hat{y}$

One of the nice features of regression is that it provides a model which allows for prediction - one obvious question about any prediction is how confident are we about it, and what is the range of plausible values it might also take?

First we want to rewrite our model that it is in terms of just b_1:

$$\begin{aligned}\hat{y}_\star &= b_0 + b_1\, x_\star \\ &= (\bar{y} - b_1\, \bar{x}) + b_1\, x_\star \\ &= \bar{y} + b_1(x_\star - \bar{x})\end{aligned}$$

To find the sampling distribution we then need to know $E(\hat{y}_\star)$ and $Var(\hat{y}_\star)$. We also need to know the distribution, which turns out to be Normal in this case.

Expected value of $\hat{y}$

We'll use the original definition of $\hat{y} = b_0 + b_1\, x_\star$ here,

$$\begin{aligned}E(\hat{y}_\star) &= E(b_0 + b_1\, x_\star) \\ &= E(b_0) + E(b_1\, x_\star) \\ &= \beta_0 + E(b_1)\, x_\star \\ &= \beta_0 + \beta_1\, x_\star\end{aligned}$$

This assumes that both b_0 and b_1 are unbiased estimators of β_0 and β_1 respectively (which is true).

Variance of $\hat{y}$

We'll use the revised definition of $\hat{y} = \bar{y} + b_1(x_\star - \bar{x})$ here,

$$\begin{aligned}Var(\hat{y}_\star) &= Var(\bar{y} + b_1(x_\star - \bar{x})) \\ &= Var(\bar{y}) + (x_\star - \bar{x})^2\; Var(b_1) \\ &= \frac{\sigma_e^2}{n} + (x_\star - \bar{x})^2\, \frac{\sigma_e^2}{\sum_i (x_i - \bar{x})^2} \\ &= \sigma_e^2 \left(\frac{1}{n} + \frac{(x_\star - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right)\end{aligned}$$

Since we don't know σ_e^2 we need to us s_e^2, which means that we need to use a t-distribution for our sampling distribution.

Confidence intervals for average values

A confidence interval for the average (expected) value of y for a given $x^\star$, is given by

$$\hat{y} \pm t^\star_{n-2}\, s_e \sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

where s_e is the standard deviation of the residuals

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Note that when $x^\star = \bar{x}$ this reduces to

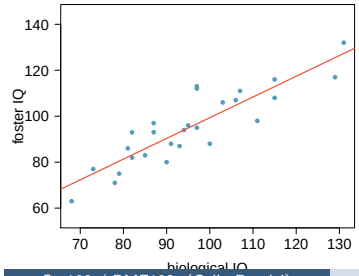
$$\hat{y} \pm t^\star_{n-2} \frac{s_e}{\sqrt{n}}$$

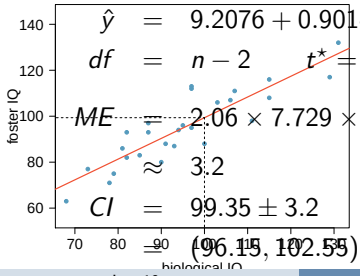
Example Calculation

Calculate a 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation is 15.74 points.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
```





$$\hat{y} = 9.2076 + 0.90144 \times 100 \approx 99.35$$

$$df = n - 2 \qquad t^\star = 2.06$$

$$ME = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}}$$

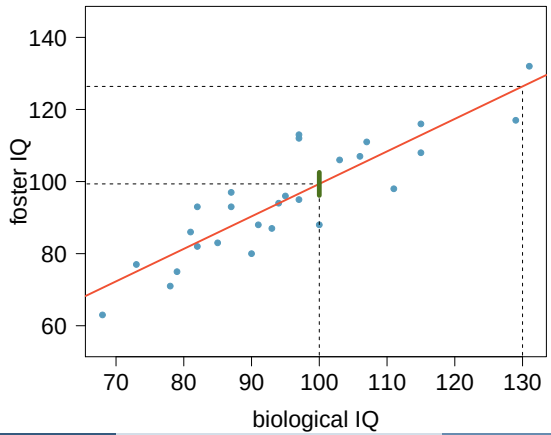
$$\approx 3.2$$

$$CI = 99.35 \pm 3.2$$

$$= (96.15, 102.55)$$

Distance from the mean

How would you expect the width of the 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 130 points ($x^\star = 130$) to compare to the previous confidence interval (where $x^\star = 100$)?



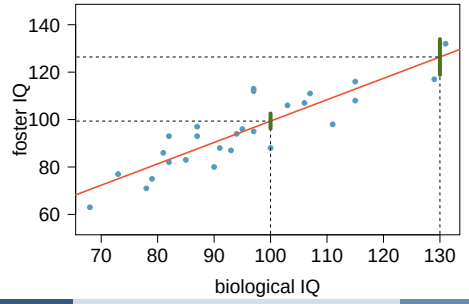
How do the confidence intervals where $x^\star = 100$ and $x^\star = 130$ compare in terms of their widths?

$x^\star = 100$

$$ME_{100} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} = 3.2$$

$x^\star = 130$

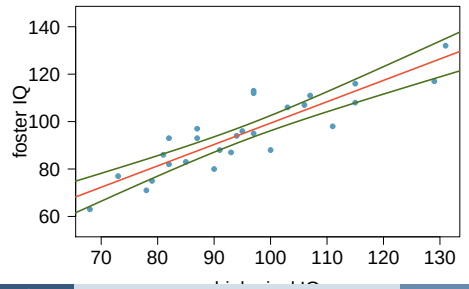
$$ME_{130} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(130 - 95.3)^2}{26 \times 15.74^2}} = 7.53$$



Recap

The width of the confidence interval for $E(y)$ increases as $x^\star$ moves away from the center.

- Conceptually: We are much more certain of our predictions at the center of the data than at the edges (and our level of certainty decreases even further when predicting outside the range of the data – extrapolation).
- Mathematically: As $(x^\star - \bar{x})^2$ term increases, the margin of error of the confidence interval increases as well.



Predicting a value, not an average

Earlier we learned how to calculate a confidence interval for average y, $E(y)$, for a given $x^\star$.

Suppose that we are not interested in the average, but instead we want to predict a future value of y for a given $x^\star$.

Would you expect there to be more uncertainty around an average or a specific predicted value?

Prediction intervals

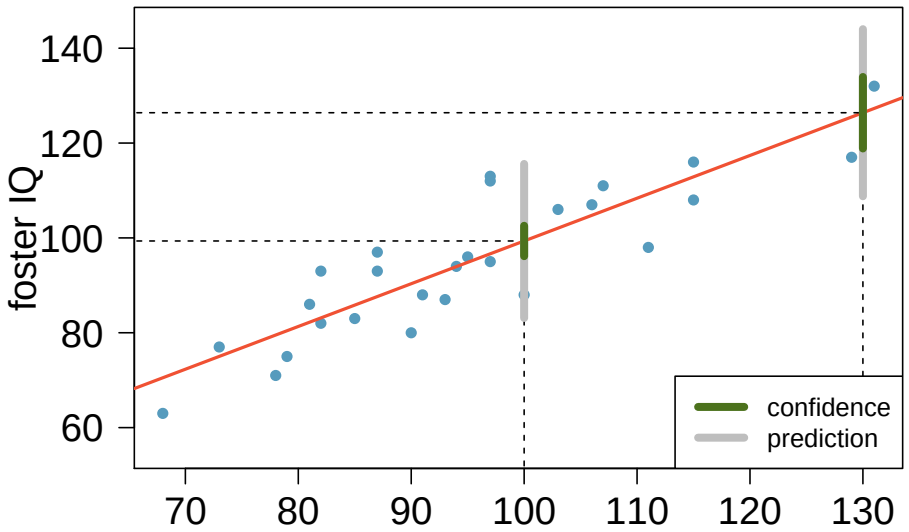
A *prediction interval* for y for a given $x^\star$ is

$$\hat{y} \pm t^\star_{n-2}\, s_e \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

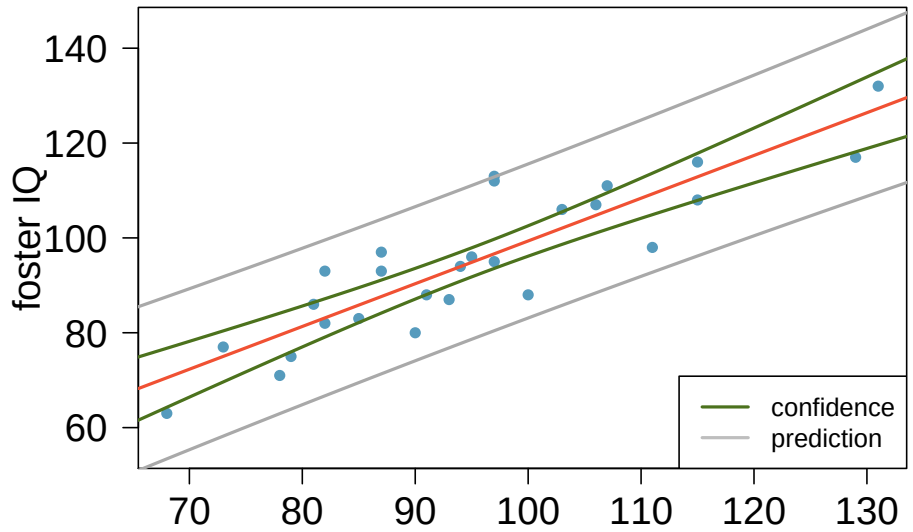
where s is the standard deviation of the residuals.

- The formula is very similar, except the variability is higher since there is a 1 added in the formula.
- Prediction level: If we repeat the study of obtaining a regression data set many times, each time forming a XX% prediction interval at $x^\star$, and wait to see what the future value of y is at $x^\star$, then roughly XX% of the prediction intervals will contain the corresponding actual value of y.

Confidence interval for $E(y)$ vs. prediction interval for y



CI for $E(y)$ vs. PI for y



CI for $E(y)$ vs. PI for y - differences

- A prediction interval is similar in spirit to a confidence interval, except that
  - the prediction interval is designed to cover a "moving target", the random future value of y
  - the confidence interval is designed to cover the "fixed target", the average (expected) value of y, $E(y)$,
- Although both are centered at $\hat{y}$, the prediction interval is wider than the confidence interval, for a given $x^\star$ and confidence level. This makes sense, since
  - the prediction interval must take account of the tendency of y to fluctuate from its mean value
  - the confidence interval simply needs to account for the uncertainty in estimating the mean value.

CI for $E(y)$ vs. PI for y - similarities

- For a given data set, the error in estimating $E(y)$ and $\hat{y}$ grows as $x^\star$ moves away from $\bar{x}$. Thus, the further $x^\star$ is from $\bar{x}$, the wider the confidence and prediction intervals will be.
- If any of the conditions underlying the model are violated, then the confidence intervals and prediction intervals may be invalid as well. This is why it's so important to check the conditions by examining the residuals, etc.