Uncertainty of predictions

Regression models are useful for making predictions for new observations not included in the original dataset.

If the model is good, the predictions should be close to the true value of the response variable for this observation, however it may not be exact, i.e. \( \hat{y} \) might be different than \( y \).

With any prediction we can (and should) also report a measure of uncertainty of the prediction:

- Use a confidence interval for the uncertainty around the expected value of predictions (average of a group of predictions) – e.g. predict the average final exam score of a group of students who scored the same on the midterm.
- Use a prediction interval for the uncertainty around a single prediction – e.g. predict the final exam score of one student with a given midterm score.

Confidence intervals for average values

A confidence interval for the average (expected) value of \( y \), \( E(y) \), for a given \( x^* \), is

\[
\hat{y} \pm t_{n-2}\frac{s}{\sqrt{n}} \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2} \right)
\]

where \( s \) is the standard deviation of the residuals, calculated as \( \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}} \).
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 of 15.74 points.

\[
\hat{\mu} = 9.2076 + 0.90144 \times 100 \approx 99.35 \\
\text{df} = n - 2 \quad t^* = 2.06 \\
\text{ME} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} \\
\approx 3.2 \\
\text{CI} = 99.35 \pm 3.2 \\
= (96.15, 102.55)
\]

### Confidence interval for a prediction – in R

```r
# load data
install.packages("faraway") # dataset can be found in this package
data(twins)

# fit model
m = lm(Foster ~ Biological, data = twins)

# create a new data frame for the new observation
newdata = data.frame(Biological = 100)

# calculate a prediction
# and a confidence interval for the prediction
predict(m, newdata, interval = "confidence")
```

<table>
<thead>
<tr>
<th>fit</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.3512</td>
<td>96.14866</td>
<td>102.5537</td>
</tr>
</tbody>
</table>

### Clicker question

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 \(x^* = 130\) to compare to the previous confidence interval (where \(x^* = 100\))?

(a) wider 
(b) narrower
(c) same width
(d) cannot tell

\[
x^* = 100 \quad ME_{100} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} = 3.2 \\
\]

\[
x^* = 130 \quad ME_{130} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(130 - 95.3)^2}{26 \times 15.74^2}} = 7.53
\]
Recap

The width of the confidence interval for $E(y)$ increases as $x^*$ 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 $(x^* - \bar{x})^2$ term increases, the margin of error of the confidence interval increases as well.

Prediction intervals for specific predicted values

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

$$\hat{y} \pm t_{n-2}^* s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

where $s$ is the standard deviation of the residuals.

- The formula is very similar, except the variability is higher since there is an added 1 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 $x^*$, and wait to see what the future value of $y$ is at $x^*$, then roughly XX% of the prediction intervals will contain the corresponding actual value of $y$.

Clicker question

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

Suppose we’re not interested in the average, but instead we want to predict a future value of $y$ for a given $x^*$.

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

(a) more uncertainty around an average
(b) more uncertainty around a specific predicted value
(c) equal uncertainty around both values
(d) cannot tell

Application exercise:

Prediction interval

Calculate a 95% prediction 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) | 9.26760 | 9.29990 | 0.999 | 0.332 |
| bioIQ | 0.96144 | 0.89633 | 9.358 | 1.2e-09 |

Residual standard error: 7.729 on 25 degrees of freedom
### Confidence interval for a prediction – in R

```r
# calculate a prediction
# and a confidence interval for the prediction
predict(m, newdata, interval = "prediction")
```

```
fitted  lwr    upr
99.3512  83.11356  115.5888
```
For a given data set, the error in estimating $E(y)$ and $\hat{y}$ grows as $x^*$ moves away from $\bar{x}$. Thus, the further $x^*$ 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.

In the case of multiple linear regression (regression with many predictors), confidence and prediction intervals for a new prediction works exactly the same way.

However the formulas are much more complicated since we no longer have just one $x$, but instead many $x$s.

For confidence and prediction intervals for MLR we will focus on the concepts and leave the calculations up to R.

### Prediction of evaluation my evaluation score

```r
# fit a model
m = lm(score ~ rank + gender + language +
       cls_perc_eval + cls_students, data = evals)

# create a data frame with the new observation (mine)
prof = data.frame(rank = "teaching", gender = "female",
                  language = "english", cls_perc_eval = 90,
                  cls_students = 100)

# prediction interval
predict(m, prof, interval = "prediction")
# fit lwr upr
4.337951 3.301877 5.374026

# confidence interval
predict(m, prof, interval = "confidence")
# fit lwr upr
4.337951 4.20273 4.473172
```

### Truck prices

The scatterplot below shows the relationship between year and price of a random sample of 43 pickup trucks. Describe the relationship between these two variables.

![Scatterplot](http://faculty.chicagobooth.edu/robert.gramacy/teaching.html)
Remove unusual observations

Let’s remove trucks older than 20 years, and only focus on trucks made in 1992 or later.

Now what can you say about the relationship?

Truck prices - linear model?

Model:
\[ \hat{\text{price}} = b_0 + b_1 \text{ year} \]

The linear model doesn’t appear to be a good fit since the residuals have non-constant variance.

Truck prices - log transform of the response variable

Model:
\[ \log(\text{price}) = b_0 + b_1 \text{ year} \]

We applied a log transformation to the response variable. The relationship now seems linear, and the residuals no longer have non-constant variance.

Interpreting models with log transformation

For each additional year the car is newer (for each year decrease in car's age) we would expect the log price of the car to increase on average by 0.14 log dollars.

which is not very useful...
Subtraction and logs: \( \log(a) - \log(b) = \log\left(\frac{a}{b}\right) \)

Natural logarithm: \( e^{\log(x)} = x \)

We can these identities to "undo" the log transformation

Interpreting models with log transformation (cont.)

The slope coefficient for the log transformed model is 0.14, meaning the log price difference between cars that are one year apart is predicted to be 0.14 log dollars.

\[
\log\left(\frac{\text{price at year } x + 1}{\text{price at year } x}\right) = 0.14 \\
e^{\log\left(\frac{\text{price at year } x + 1}{\text{price at year } x}\right)} = e^{0.14} \\
\frac{\text{price at year } x + 1}{\text{price at year } x} = 1.15
\]

For each additional year the car is newer (for each year decrease in car's age) we would expect the price of the car to increase on average by a factor of 1.15.

Recap: dealing with non-constant variance

- Non-constant variance is one of the most common model violations, however it is usually fixable by transforming the response \( y \) variable
- The most common variance stabilizing transform is the log transformation: \( \log(y) \), especially useful when the response variable is (extremely) right skewed.
- When using a log transformation on the response variable the interpretation of the slope changes:
  - For each unit increase in \( x \), \( y \) is expected on average to decrease/increase by a factor of \( e^{b_1} \).
- Another useful transformation is the square root: \( \sqrt{y} \), especially useful when the response variable is counts.
- These transformations may also be useful when the relationship is non-linear, but in those cases a polynomial regression may also be needed.