IDS 702: Module 1.5

Checking main regression assumptions

Dr. Olanrewaju Michael Akande

Assumptions for MLR

- Inference (CIs, p-values, or predictions) can only be meaningful when the regression assumptions are plausible.
- The main assumptions follow directly from SLR.
- They are:

1. Linearity (response variable vs. each predictor)
2. Independence of errors (really just independence in the observations)
3. Equal variance (in the error term $\Rightarrow$ in the response variable)
4. Normality (in the error term $\Rightarrow$ in the response variable)

First, think about validity of the model

- Validity is about whether the data and model are even suitable for answering the research question.
- To quote the Gelman and Hill book,

  "Optimally, this means that the outcome measure should accurately reflect the phenomenon of interest, the model should include all relevant predictors, and the model should generalize to cases to which it will be applied.

  For example, a model of earnings will not necessarily tell you about pattern of total assets, neither will a model of test scores necessarily tell you about child intelligence or cognitive development."

- You should always keep in mind the types of questions you can and cannot answer reliably from both the data and model.

Checking assumptions

- Checking all four assumptions usually requires examining the residuals after model fitting.
- For linearity, a plot of the residuals versus each predictor will often do.

  What should we look out for in such a plot?
- Note that the residuals contain information about the response variable variable y that has not been explained by the covariates in the model.
- Thus, when we plot the residuals against each predictor, we should NOT expect to see any pattern.
- Some pattern in any of the plots is usually an indication of a relationship (often nonlinear) between y and that predictor, which has not been captured yet in the model.

Checking assumptions

- To check both independence of the errors and the equal variance assumption, usually a plot of the residuals versus the fitted values will do.
- Can also do a plot of the residuals versus the indexes of the observations.
- The points in the plot should look random (for independence) and be "roughly" equally spread out around zero (for equal variance).
- To check normality, it is often sufficient to look at a qq-plot (quantile-quantile plot) which compares the distribution of the standardized residuals to the theoretical quantiles of a standard normal distribution.
- Clustering of the points around the 45 degree line of the qq-plot usually implies the normality assumption is not violated.
- One can also look at a histogram of the residuals, but it is much harder to judge deviations from normality through histograms.
- Can also do a test of normality: Shapiro–Wilk, Kolmogorov–Smirnov, etc.

Checking assumptions

Let's revisit the same data from last class. Recall the model we fit to the data and the results:

$$\text{bsal}_i = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{senior}_i + \beta_3 \text{age}_i + \beta_4 \text{educ}_i + \beta_5 \text{exper}_i + \epsilon_i$$

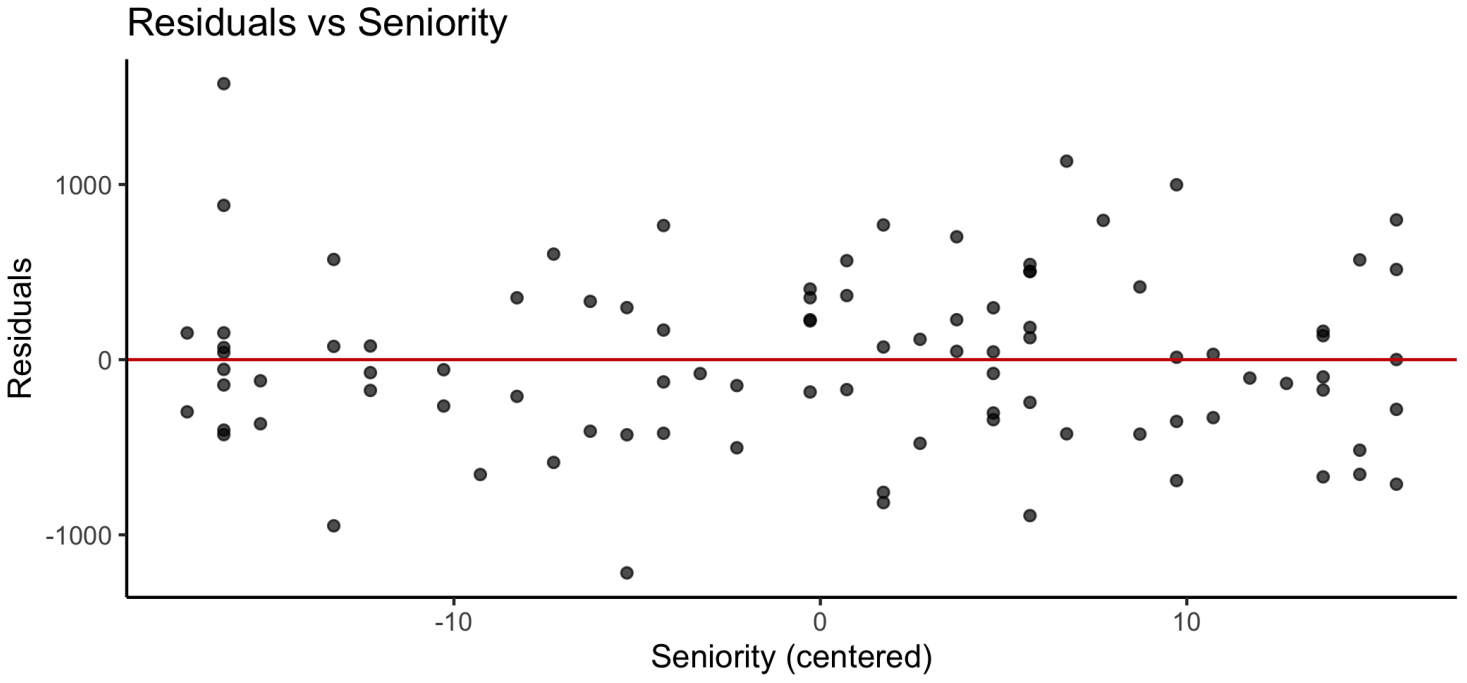
```
regwagec <- lm(bsal~ sex + seniorc + agec + educc + experc, data= wages)
summary(regwagec)
```

```
##
## Call:
## lm(formula = bsal ~ sex + seniorc + agec + educc + experc, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1217.36  -342.83   -55.61   297.10  1575.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5924.0072    99.6588  59.443  < 2e-16
## sexFemale   -767.9127   128.9700  -5.954 5.39e-08
## seniorc      -22.5823     5.2957  -4.264 5.08e-05
## agec           0.6310     0.7207   0.876 0.383692
## educc         92.3060    24.8635   3.713 0.000361
## experc         0.5006     1.0553   0.474 0.636388
##
## Residual standard error: 508.1 on 87 degrees of freedom
## Multiple R-squared:  0.5152,    Adjusted R-squared:  0.4873
## F-statistic: 18.49 on 5 and 87 DF,  p-value: 1.811e-12
```

Checking linearity

Now, let's plot the residuals against each predictor. First, let's look at senior.

```
ggplot(wages,aes(x=seniorc, y=regwagec$residual)) +
  geom_point(alpha = .7) +  geom_hline(yintercept=0,col="red3") + theme_classic() +
  labs(title="Residuals vs Seniority",x="Seniority (centered)",y="Residuals")
```

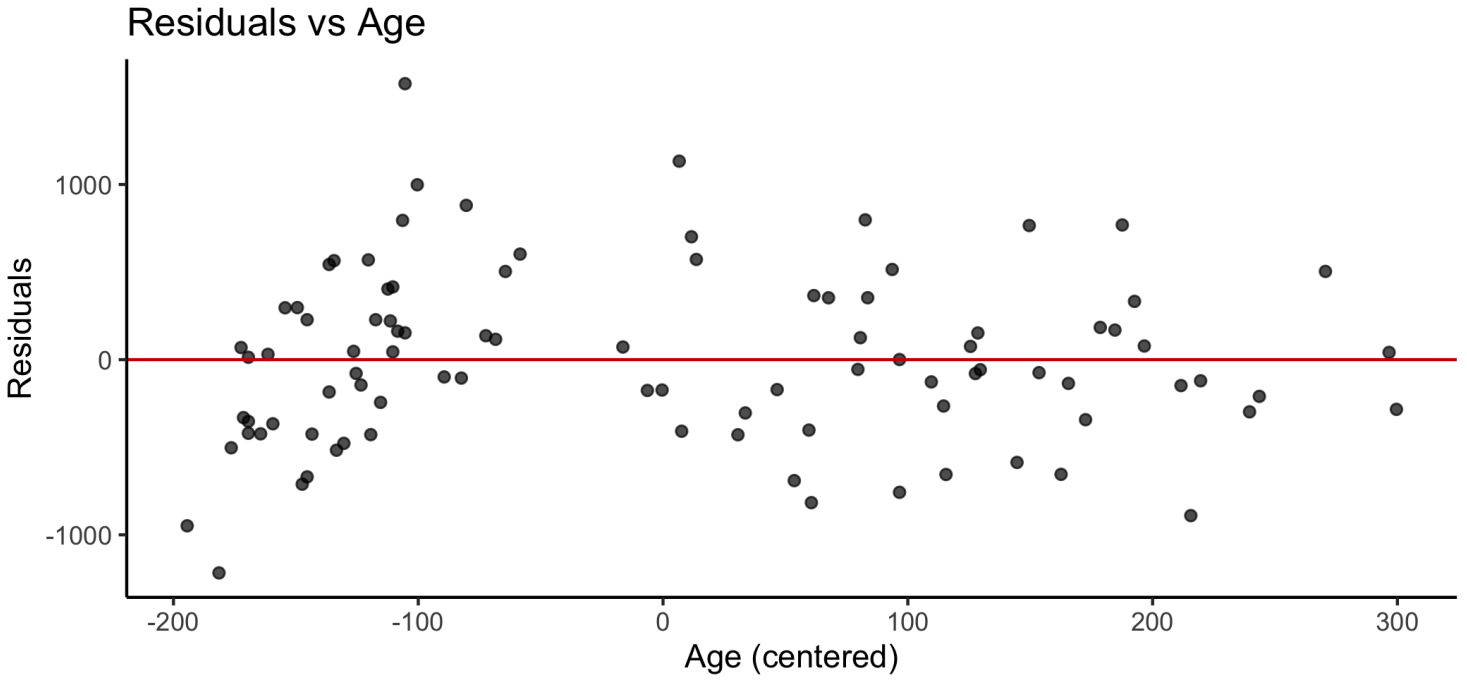


Are there any clear violations of the linearity assumption?

Checking linearity

Next, let's look at age.

```
ggplot(wages,aes(x=agec, y=regwagec$residual)) +
  geom_point(alpha = .7) +  geom_hline(yintercept=0,col="red3") + theme_classic() +
  labs(title="Residuals vs Age",x="Age (centered)",y="Residuals")
```

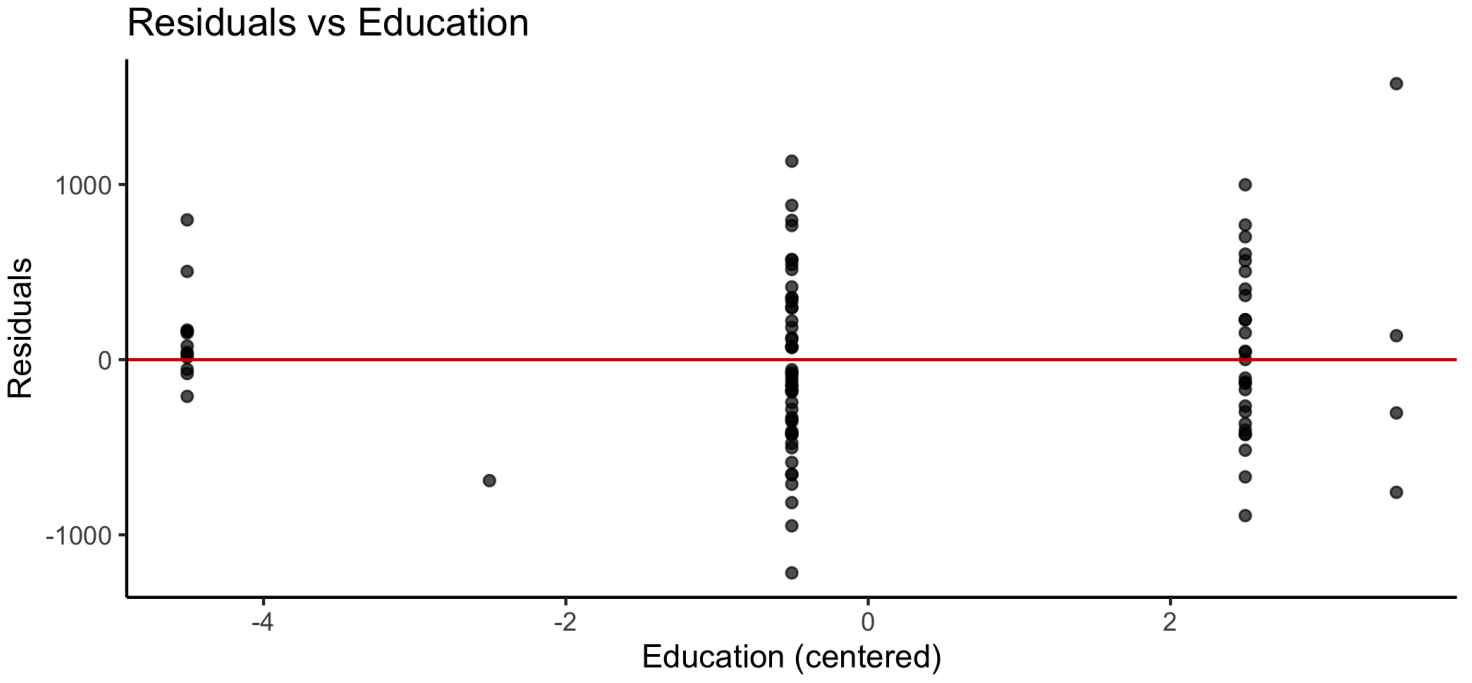


Are there any clear violations of the linearity assumption?

Checking linearity

Next, let's look at educ.

```
ggplot(wages,aes(x=educc, y=regwagec$residual)) +
  geom_point(alpha = .7) +  geom_hline(yintercept=0,col="red3") + theme_classic() +
  labs(title="Residuals vs Education",x="Education (centered)",y="Residuals")
```

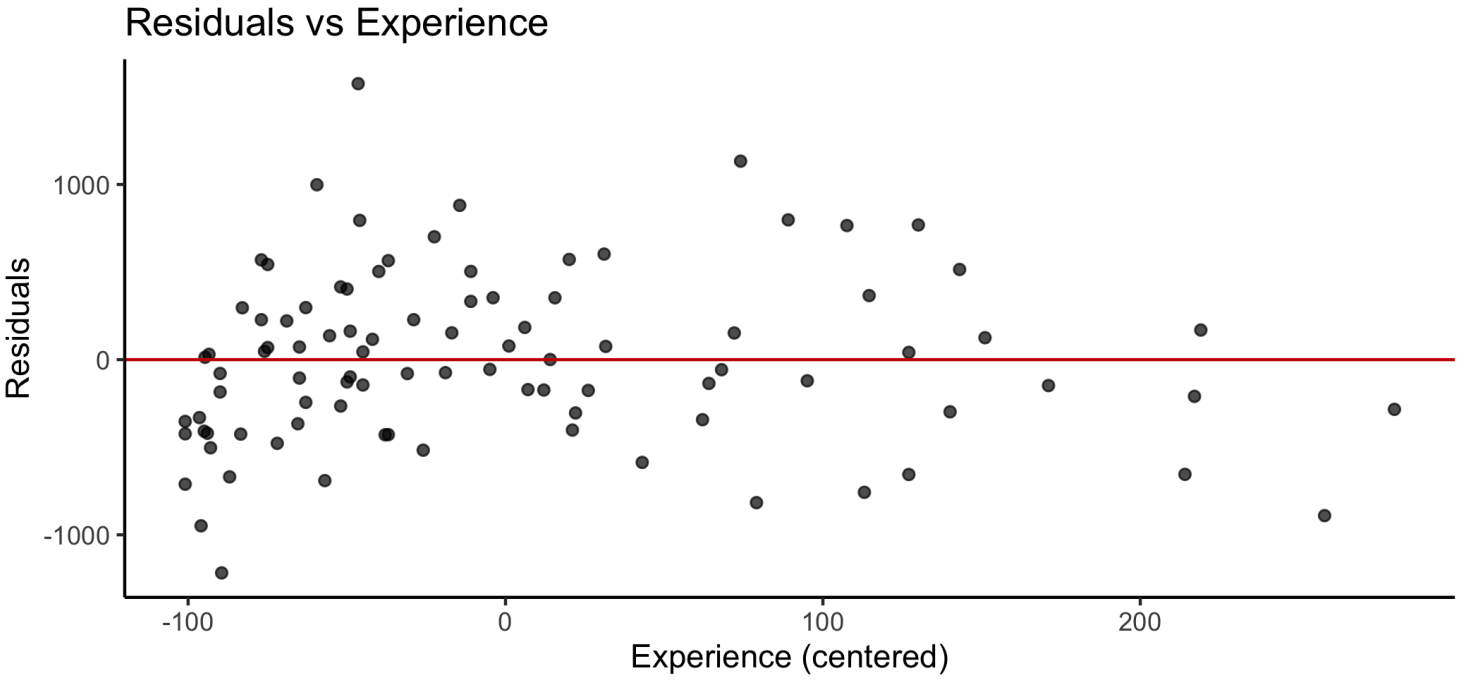


Education is an odd one to visualize this way. We will revisit this issue later.

Checking linearity

Next, let's look at exper.

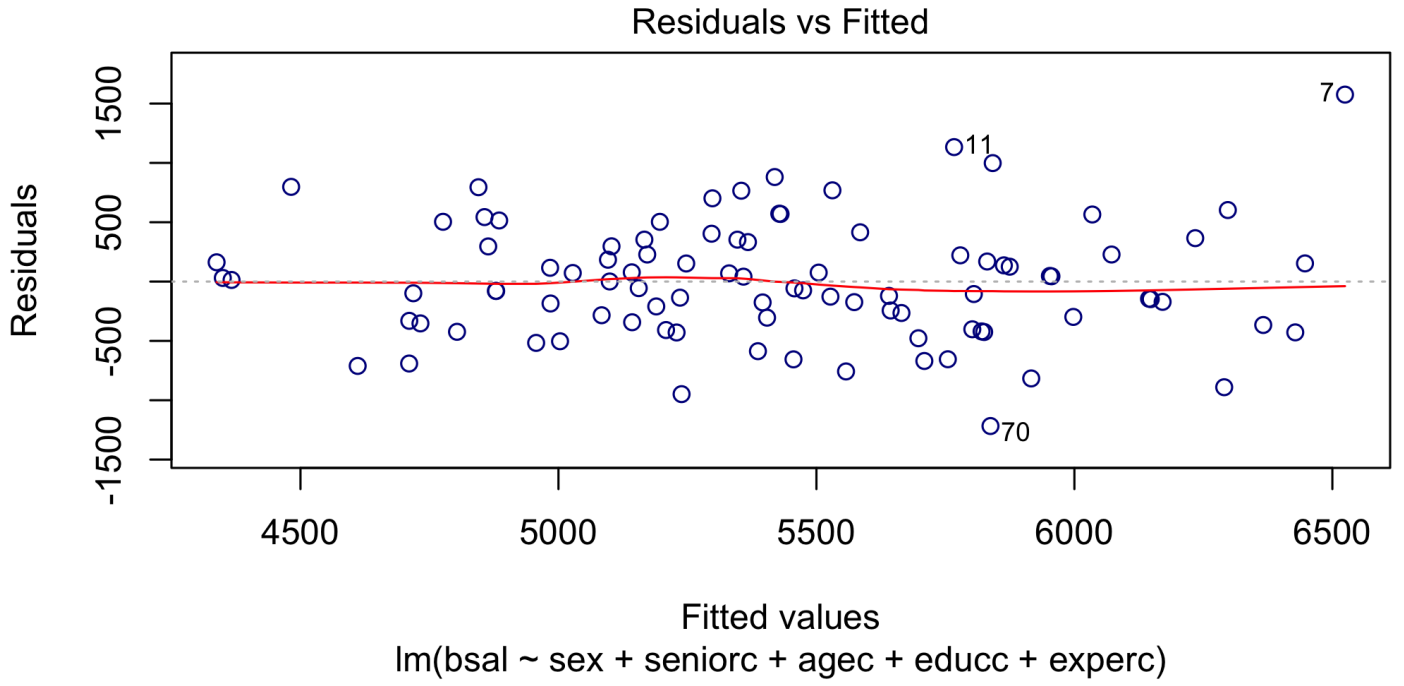
```
ggplot(wages,aes(x=experc, y=regwagec$residual)) +
  geom_point(alpha = .7) +  geom_hline(yintercept=0,col="red3") + theme_classic() +
  labs(title="Residuals vs Experience",x="Experience (centered)",y="Residuals")
```



Are there any clear violations of the linearity assumption?

Checking independence and equal variance

```
plot(regwagec,which=1,col=c("blue4"))
```

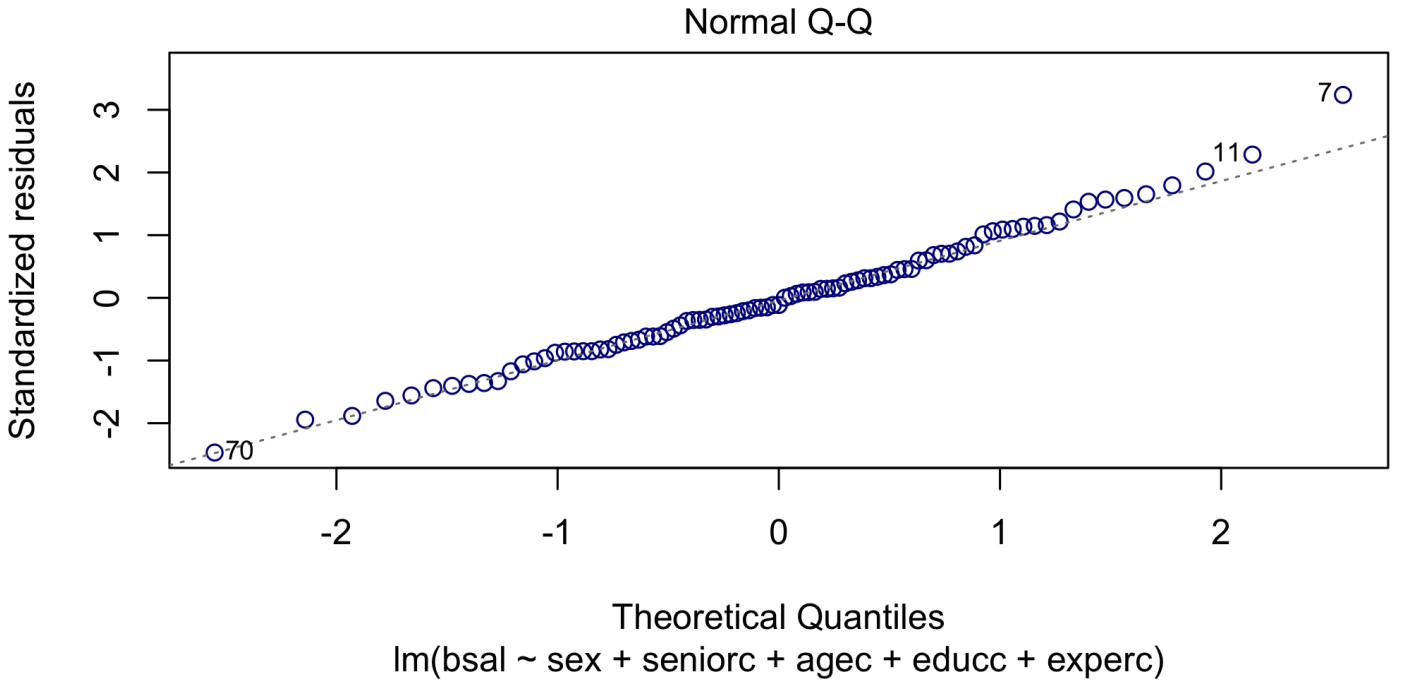


Are there any clear violations of the independence and equal variance assumptions?

IDS 702

Checking normality

```
plot(regwagec,which=2,col=c("blue4"))
```



Are there any clear violations of the normality assumption?

Takeaways from residual plots

- Looks like we may have to worry about the assumption of linearity being violated for age and experience.
- There appears to be some quadratic trend for both variables and possible non-constant variance, so let's improve the model by adding the squared term for each variable.
- Let's add the squared terms of the centered age and centered experience predictors to the dataset and refit the model.

```
wages$agec2 <- wages$agec^2
wages$experc2 <- wages$experc^2
```

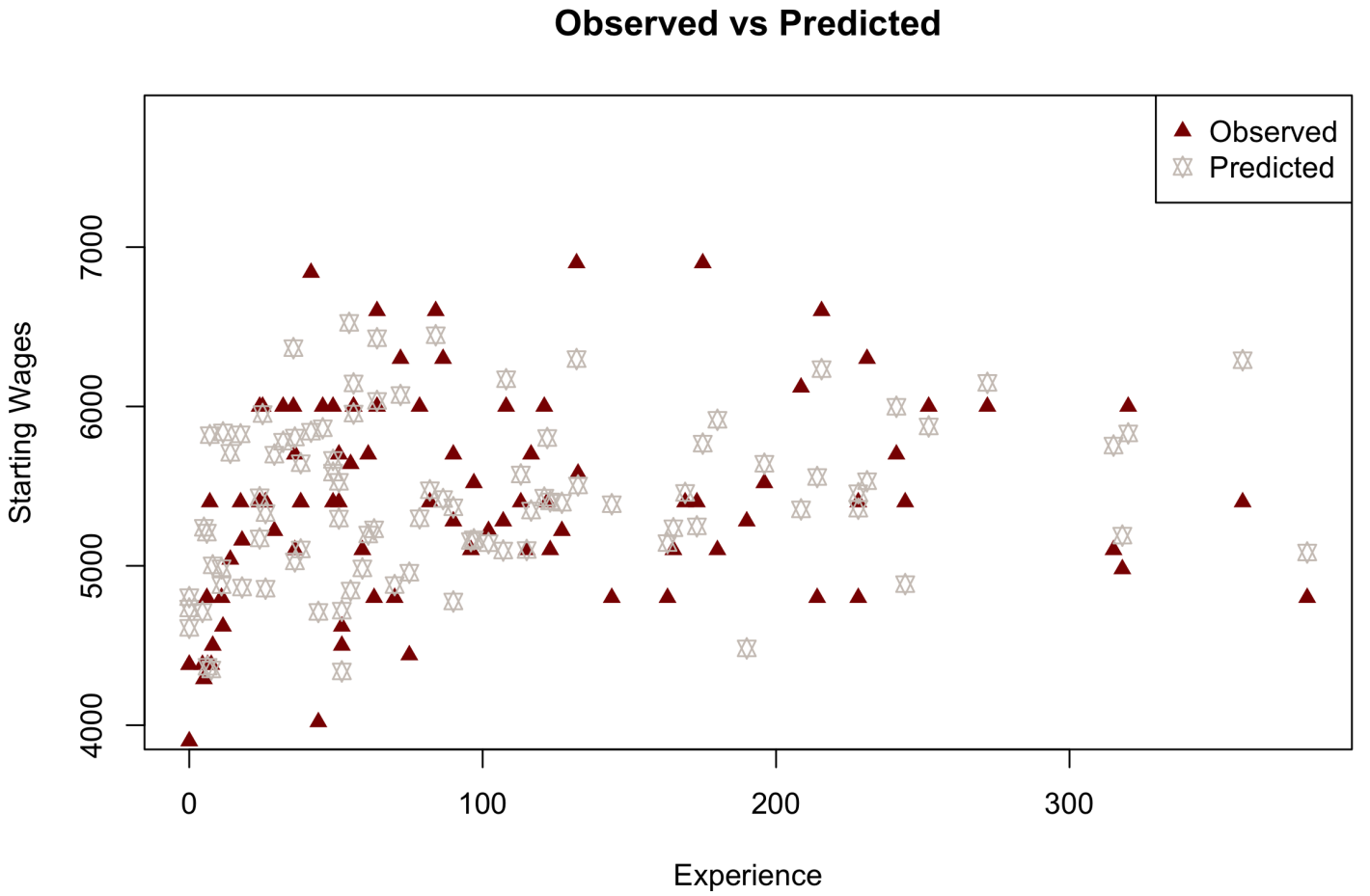
Re-fitting the model

```
regwagecsquares <- lm(bsal~sex+seniorc+agec+agec2+educc+experc+experc2,data=wages)
summary(regwagecsquares)
```

```
##
## Call:
## lm(formula = bsal ~ sex + seniorc + agec + agec2 + educc + experc +
##     experc2, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1086.84  -267.84    -8.71   304.92  1642.44
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.098e+03  1.123e+02  54.313  < 2e-16
## sexFemale   -7.684e+02  1.211e+02  -6.343 1.04e-08
## seniorc     -1.764e+01  5.265e+00  -3.351  0.00120
## agec        -3.473e-01  7.814e-01  -0.444  0.65783
## agec2        7.195e-04  4.045e-03   0.178  0.85925
## educc        7.561e+01  2.406e+01   3.142  0.00231
## experc       4.035e+00  1.479e+00   2.729  0.00772
## experc2     -2.298e-02  7.592e-03  -3.027  0.00326
##
## Residual standard error: 477 on 85 degrees of freedom
## Multiple R-squared:  0.5825,    Adjusted R-squared:  0.5481
## F-statistic: 16.94 on 7 and 85 DF,  p-value: 8.011e-14
```

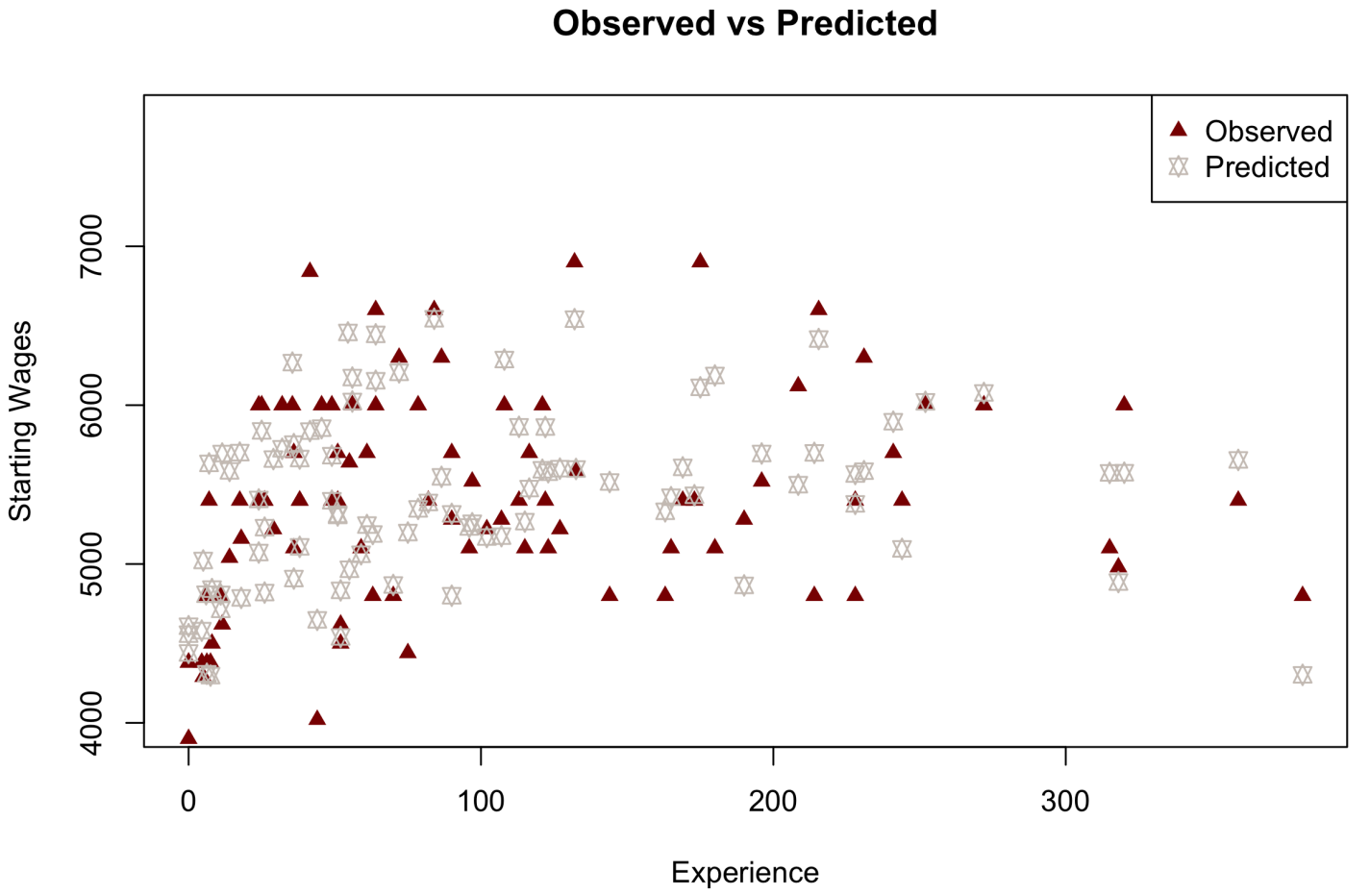
How well does the model fit the data.

The first model: NO quadratic term



How well does the model fit the data.

Now the latest model: add quadratic term



Interpreting the new model

Clearly, experc2 is an important predictor given all other predictors. However, it can be tough to interpret its effect using the coefficient. Instead, let's visualize the effect of changing experience.

```
#First, make the 20 values of experience that you want to examine
newexper <- seq(from=0,to=400,by=5)
newexperc <- newexper - mean(wages$exper)
newexperc2 <- newexperc^2
newdata <- data.frame(matrix(0, nrow=length(newexper), ncol=7))
names(newdata) <- c("sex","seniorc","agec","agec2","educc","experc","experc2")
newdata$experc <- newexperc; newdata$experc2 <- newexperc2; newdata$sex <- "Male"
#Since we use mean-centered predictors, the rows in the new dataset correspond to
#people with average values of seniority, age, and education.
preds_male <- predict(regwagecsquares,newdata,interval="confidence"); preds_male[1:3,]
```

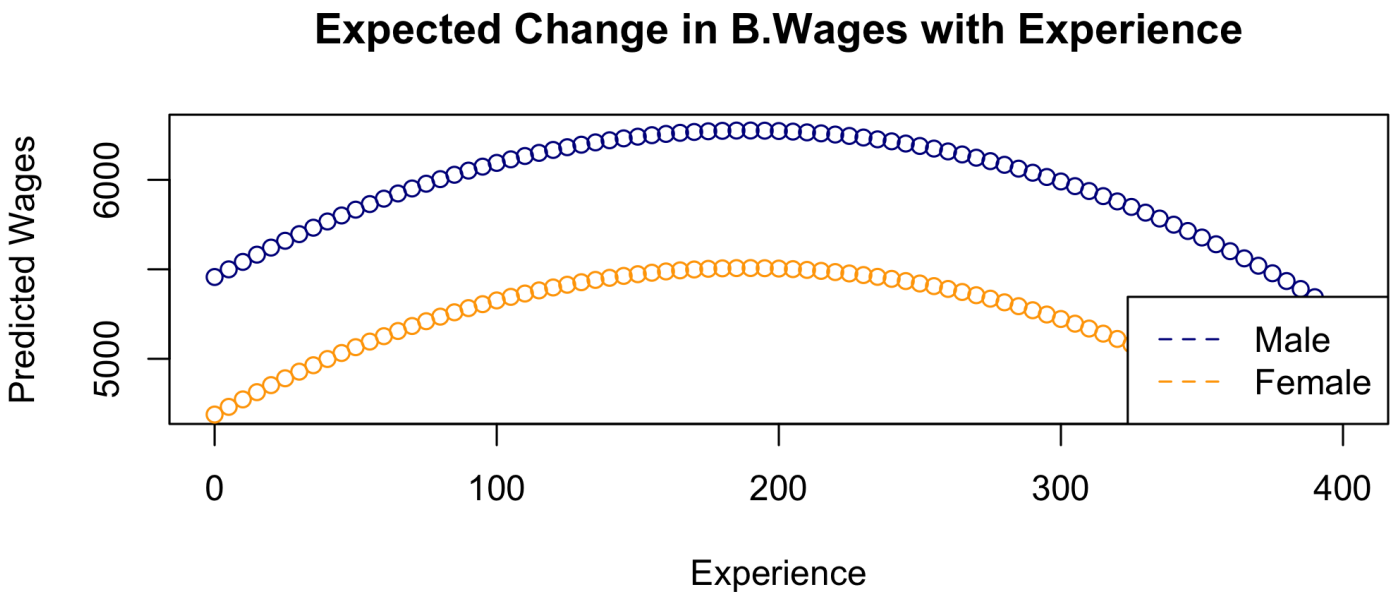
```
##        fit      lwr      upr
## 1 5457.022 4983.617 5930.426
## 2 5499.822 5049.929 5949.715
## 3 5541.473 5114.091 5968.854
```

```
newdata$sex <- "Female";
preds_female <- predict(regwagecsquares,newdata,interval="confidence");
preds_female[1:3,]
```

```
##        fit      lwr      upr
## 1 4688.582 4269.895 5107.269
## 2 4731.382 4337.545 5125.219
## 3 4773.033 4403.057 5143.009
```

Interpreting the new model

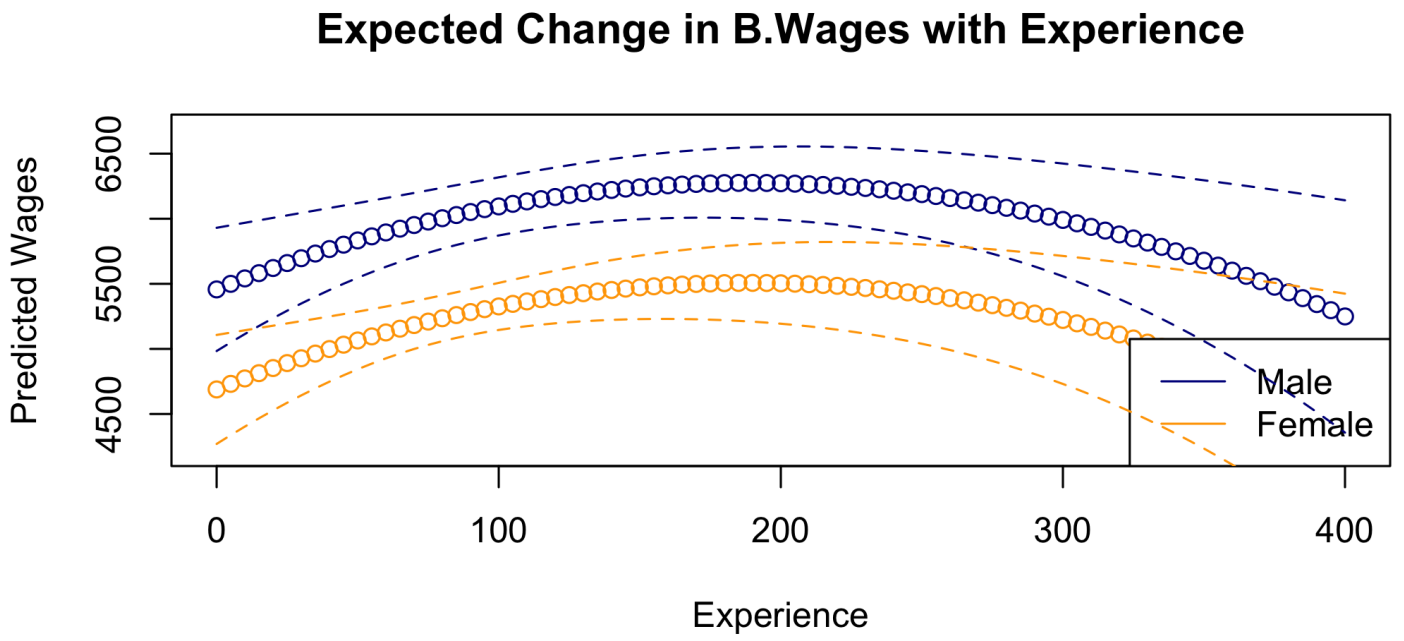
```
plot(y=preds_male[,"fit"],x=newexper,xlab="Experience",ylab="Predicted Wages",
     main="Expected Change in B.Wages with Experience",col="darkblue",ylim=c(4700,6300))
points(y=preds_female[,"fit"], x=newexper,col="orange")
legend("bottomright",c("Male","Female"),col=c("darkblue","orange"),lty=c(2,2))
#Remember that this is with average values of other predictors.
```



Why do we have the exact same trend for male and female?

INTERPRETING THE NEW MODEL

```
#if you want to get the 95% confidence bands on the plot as well, you can do the following
plot(y=preds_male[,"fit"], x=newexper, xlab="Experience", ylab="Predicted Wages",
     main="Expected Change in B.Wages with Experience",col="darkblue",ylim=c(4200,6700))
points(y=preds_female[,"fit"], x=newexper,col="orange")
legend("bottomright",c("Male","Female"),col=c("darkblue","orange"),lty=c(1,1))
lines(y=preds_male[,"lwr"],x=newexper,col="darkblue",lty=2)
lines(y=preds_male[,"upr"],x=newexper,col="darkblue",lty=2)
lines(y=preds_female[,"lwr"],x=newexper,col="orange",lty=2)
lines(y=preds_female[,"upr"],x=newexper,col="orange",lty=2)
```



Final notes

- Generally it is a good idea to start with exploratory data analysis (which we did a bit of in the last class) rather than jumping right into modeling.
- After fitting your model, model assessment is A MUST!
- In this class and outside of it, you MUST always assess your models!
- Later, we will continue with model assessment by exploring leverage, influence, and standardized residuals.
- We will also dive into model validation.

What's next?

Move on to the readings for the next module!