Agenda

1. More inference for multiple regression

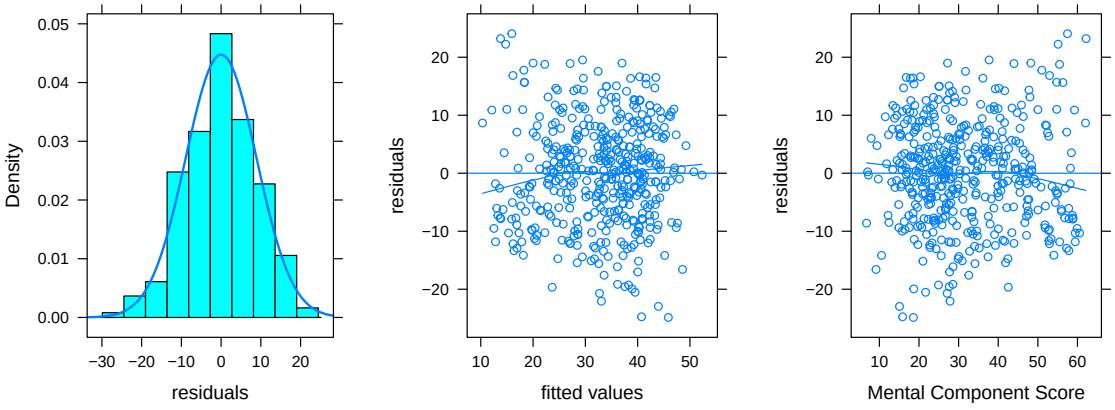
Case study: Predictors of depressive symptoms In the HELP (Health Evaluation and Linkage to Primary Care) study, investigators were interested in determining predictors of severe depressive symptoms (measured by the Center for Epidemiologic Studies - Depression scale, aka `cesd`) amongst a cohort enrolled at a substance abuse treatment facility. These includes `substance` of abuse (alcohol, cocaine, or heroin), `mcs` (a measure of mental well-being), gender and housing status (housed or homeless). Consider the following multiple regression model.

```
require(mosaic)
fm <- lm(cesd ~ substance + mcs + sex + homeless, data = HELPrct)
msummary(fm)

##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      57.77942    1.46640  39.402  < 2e-16 ***
## substancecocaine -3.54056    1.01013  -3.505 0.000503 ***
## substanceheroin  -1.68181    1.07310  -1.567 0.117766
## mcs              -0.64073    0.03377 -18.971  < 2e-16 ***
## sexmale          -3.32387    1.00749  -3.299 0.001047 **
## homelesshoused   -0.83270    0.86864  -0.959 0.338265
##
## Residual standard error: 8.973 on 447 degrees of freedom
## Multiple R-squared:  0.4915,Adjusted R-squared:  0.4859
## F-statistic: 86.43 on 5 and 447 DF,  p-value: < 2.2e-16

confint(fm)

##                       2.5 %     97.5 %
## (Intercept)      54.8975311 60.6613125
## substancecocaine -5.5257585 -1.5553529
## substanceheroin  -3.7907660  0.4271435
## mcs              -0.7071036 -0.5743498
## sexmale          -5.3038731 -1.3438759
## homelesshoused   -2.5398149  0.8744190
```



1. Write out the linear model

2. Calculate the predicted CESD for a female homeless cocaine-involved subject with an MCS score of 20.

3. Interpret the 95% confidence interval for the `substancecocaine` coefficient

4. Make a conclusion and summarize the results of a test of the `homeless` parameter

5. Report and interpret the R^2 (coefficient of determination) for this model

6. What do we conclude about the distribution of the residuals?

7. What do we conclude about the relationship between the fitted values and the residuals?

8. What do we conclude about the relationship between the MCS score and the residuals?

9. What other things can we learn from the residual diagnostics?

10. Which observations should we flag for further study?

```
help_mod <- broom::augment(fm)
help_mod %>%
  slice(c(40, 351, 433, 450))
```

Case Study: Gestation redux The Child Health and Development Studies investigate a range of topics. One study, in particular, considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. The goal is to model the weight of the infants (`bwt`, in ounces) using variables including length of pregnancy in days (`gestation`), mother's age in years (`age`), mother's height in inches (`height`), whether the child was the first born (`parity`), mother's pregnancy weight in pounds (`weight`), and whether the mother was a smoker (`smoke`). The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.

```
require(openintro)
mod <- lm(bwt ~ gestation + age + height + weight + parity + smoke, data = babies)
coef(mod)

##   (Intercept)     gestation           age        height        weight
## -80.410853396   0.443978339  -0.008950305   1.154020364   0.050165027
##        parity         smoke
##  -3.327199613  -8.400733484
```

1. The coefficient for `parity` is different than if you fit a linear model predict weight using only that variable. Why might there be a difference?

```
coef(lm(bwt ~ parity, data = babies))


## (Intercept)      parity
##  120.068404   -1.928721
```

2. Calculate the residual for the first observation in the data set.

```
head(babies, 1)


##   case bwt gestation parity age height weight smoke
## 1    1 120       284      0  27     62    100     0


# head(fitted(mod), 1)
# head(residuals(mod), 1)
```

3. The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data used to build the model is 335.94. Calculate the R^2 and the adjusted R^2. Note that there are 1236 observations in the data set, but there was missing data in 62 of those observations, so only 1174 observations were used to build the regression model.

```
var(~residuals(mod))


## [1] 249.2832


var(~bwt, data = mod$model)


## [1] 335.9402


# rsquared(mod)
```

4. This data set contains missing values. What happens to these rows?

Regression Diagnostics

- **L**inearity - **I**ndependence - **N**ormality (of residuals) - **E**qual Variance (of residuals)
- Residuals vs. Fitted Values plot (linearity and equal variance)
- Residuals vs. each numerical explanatory variable (linearity and equal variance)
- Histogram and/or QQ-plot of residuals (normality of residuals)
- Investigate outliers and influentional points
- Investigate possible multicollinearity

You can roll your own:

```
babies_mod = broom::augment(mod)
qplot(y = .resid, x = .fitted, data = babies_mod) + geom_smooth()
qplot(y = .resid, x = gestation, data = babies_mod) + geom_smooth()
qplot(y = .resid, x = age, data = babies_mod) + geom_smooth()
qplot(y = .resid, x = height, data = babies_mod) + geom_smooth()
qplot(y = .resid, x = weight, data = babies_mod) + geom_smooth()
qplot(sample = .resid, data = babies_mod, geom = "qq")
qplot(x = .resid, data = babies_mod, geom = "blank") +
  geom_histogram(aes(y = ..density..), binwidth = 4) +
  stat_function(fun = dnorm, args = c(mean = 0, sd = sd(babies_mod$.resid)), col = "tomato")
```

Or use the built-in diagnostics:

```
plot(mod)
```

Are there interesting individual observations?

```
slice(babies, c(261, 435))
```

```
##   case bwt gestation parity age height weight smoke
## 1  261 116       148      0  28     66    135     0
## 2  435 146       263      0  39     53    110     1
```

Are there strong *pairwise* correlations between any of the explanatory variables?

```
nbabies <- babies %>%
  select(gestation, age, height, weight)
cor(nbabies, use = "complete.obs")
```

```
##             gestation          age       height     weight
## gestation  1.00000000 -0.050760907  0.068582887 0.02226461
## age       -0.05076091  1.000000000 -0.005091229 0.14759648
## height     0.06858289 -0.005091229  1.000000000 0.43631252
## weight     0.02226461  0.147596481  0.436312519 1.00000000
```

```
pairs(nbabies)
```