Agenda

1. Inference for multiple regression
2. More regression diagnostics
3. Bootstrap for regression

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.

```
library(mosaic)
mod <- lm(wt ~ gestation + age + ht + wt.1 + parity + smoke, data = Gestation)
msummary(mod)

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -88.70273   14.85467  -5.971 3.12e-09 ***
## gestation     0.46317    0.03013  15.371  < 2e-16 ***
## age           0.02450    0.09681   0.253   0.8002
## ht            1.09399    0.21266   5.144 3.15e-07 ***
## wt.1          0.06067    0.02628   2.309   0.0211 *
## parity        0.56354    0.30092   1.873   0.0614 .
## smoke        -0.88334    0.52846  -1.672   0.0949 .
##
## Residual standard error: 16.35 on 1167 degrees of freedom
##   (62 observations deleted due to missingness)
## Multiple R-squared:  0.2087,Adjusted R-squared:  0.2046
## F-statistic: 51.29 on 6 and 1167 DF,  p-value: < 2.2e-16

confint(mod)

##                     2.5 %      97.5 %
## (Intercept) -1.178476e+02 -59.5578925
## gestation    4.040487e-01   0.5222936
## age         -1.654278e-01   0.2144369
## ht           6.767417e-01   1.5112356
## wt.1         9.108466e-03   0.1122317
## parity      -2.687411e-02   1.1539491
## smoke       -1.920186e+00   0.1535040
```

1. Write the equation of the regression line that includes all of the variables.

2. Interpret the slopes of `gestation` and `age` in this context.

3. Identify the null and alternative hypotheses for the 6 tests displayed above.

4. Interpret the 95% confidence interval for the `smoke` coefficient

5. 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(wt ~ parity, data = Gestation))

## (Intercept)      parity
## 119.0369557   0.2794484
```

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

```
head(Gestation, 1)

##   id plurality outcome date gestation sex  wt parity race age ed ht wt.1
## 1 15         5       1 1411       284   1 120      1    8  27  5 62  100
##   drace dage ded dht dwt marital inc smoke time number
## 1     8   31   5  65 110       1   1     0    0      0

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

7. 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] 265.8434

var(~wt, data = mod$model)

## [1] 335.9402

# rsquared(mod)
```

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

9. Interpret the R^2 (coefficient of determination) for this model

Regression Diagnostics

- **L**inearity– scatterplot (only in s.l.r.), residual vs. fitted plot
- **I**ndependence– the thinking condition
- **N**ormality (of residuals)– QQ plot or histogram of residuals
- **E**qual Variance (of residuals)– residual vs. fitted plot

- Investigate outliers and influentional points

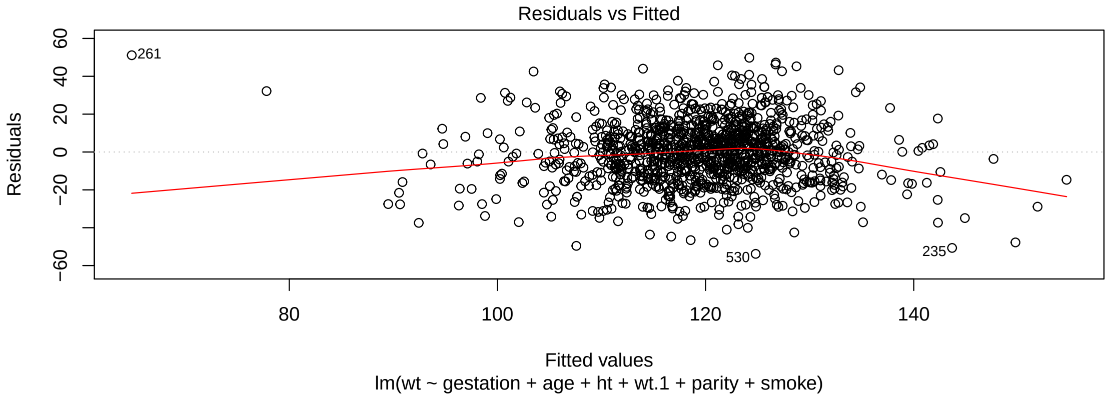
- Investigate possible multicollinearity

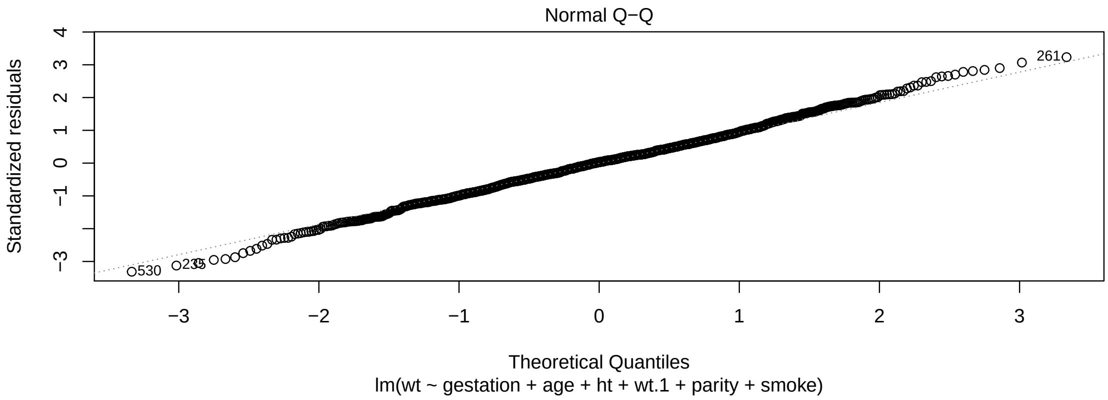
Residual analysis You can roll your own:

```
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 = ht, data = babies_mod) + geom_smooth()
qplot(y = .resid, x = wt.1, 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, which=c(1,2))
```





What do you think about the conditions for this model? Are they upheld?

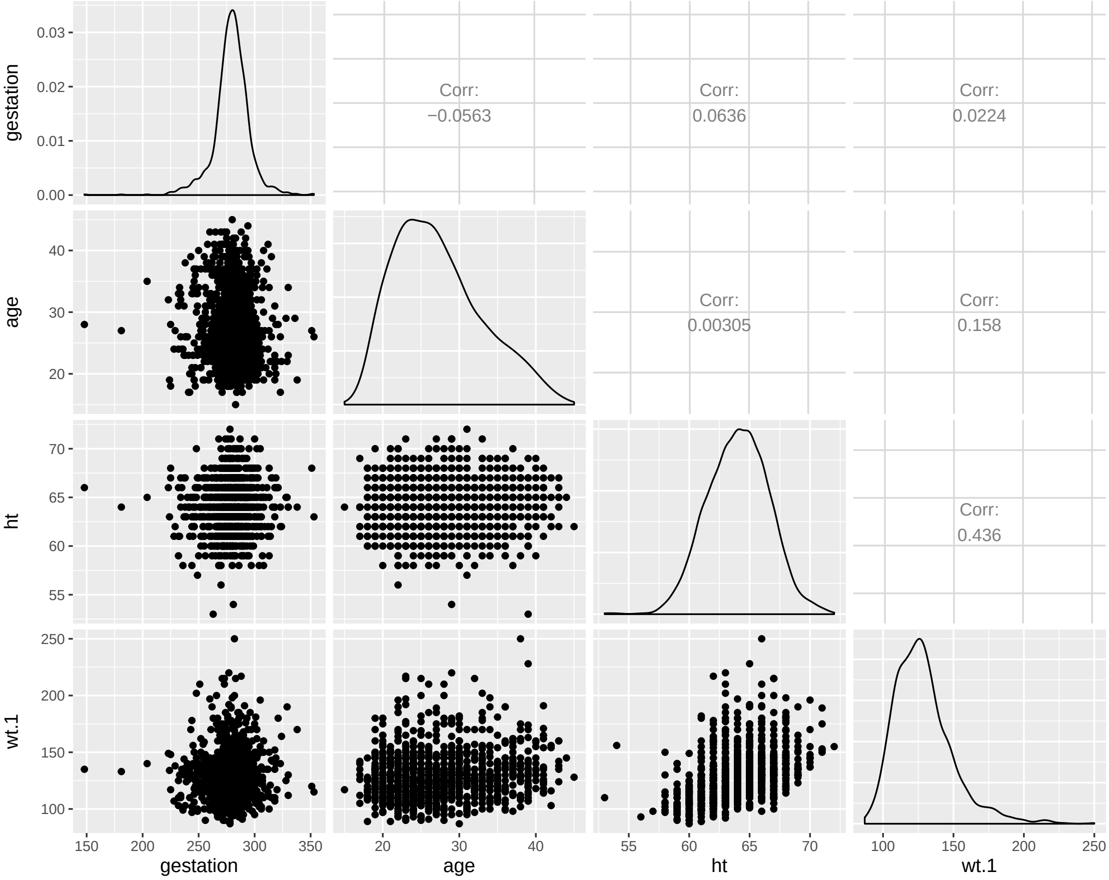
Interesting observations Are there interesting individual observations?

```
glimpse(slice(Gestation, c(261, 235)))

## Observations: 2
## Variables: 23
## $ id        <int> 4604, 3863
## $ pluralty  <int> 5, 5
## $ outcome   <int> 1, 1
## $ date      <int> 1598, 1610
## $ gestation <int> 148, 318
## $ sex       <int> 1, 1
## $ wt        <int> 116, 93
## $ parity    <int> 7, 7
## $ race      <int> 7, 7
## $ age       <int> 28, 31
## $ ed        <int> 1, 0
## $ ht        <int> 66, 66
## $ wt.1      <int> 135, 135
## $ drace     <fctr> 7, 7
## $ dage      <int> 36, 30
## $ ded       <int> 1, 4
## $ dht       <int> 68, NA
## $ dwt       <int> 155, NA
## $ marital   <int> 0, 1
## $ inc       <int> 2, NA
## $ smoke     <int> 0, 0
## $ time      <int> 0, 0
## $ number    <int> 0, 0
```

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

```
library(GGally)
nbabies <- Gestation %>%
  select(gestation, age, ht, wt.1)
ggpairs(nbabies)
```



Bootstrap for Regression Recall that a slope coefficient is an *average* or *expected* change in the response variable as a function of a unit change in that explanatory variable, holding the other explanatory variables constant. Like the sample mean, the estimated coefficient ($\hat{\beta}_1$) is a deterministic calculation based on a single sample of data, but it too has a sampling distribution. Thus, we can use the bootstrap percentile method to construct a confidence interval for it. The default confidence interval is constructed using the *t*-distribution.

```
require(mosaicData)
mod <- lm(wt ˜ age, data=Gestation)
coef(mod)

## (Intercept)         age
## 116.6834606   0.1062233

confint(mod)

##                   2.5 %     97.5 %
## (Intercept) 111.77014517 121.596776
## age          -0.07012632   0.282573
```

The bootstrap percentile method should give us a similar interval:

```
bstrap <- do(1000) * coef(lm(wt ˜ age, data = resample(Gestation)))
qdata(˜age, p = c(0.025, 0.975), data = bstrap)

##          quantile     p
## 2.5%  -0.07015013 0.025
## 97.5%  0.29114906 0.975
```