lab-8-seandenny

Sean Denny

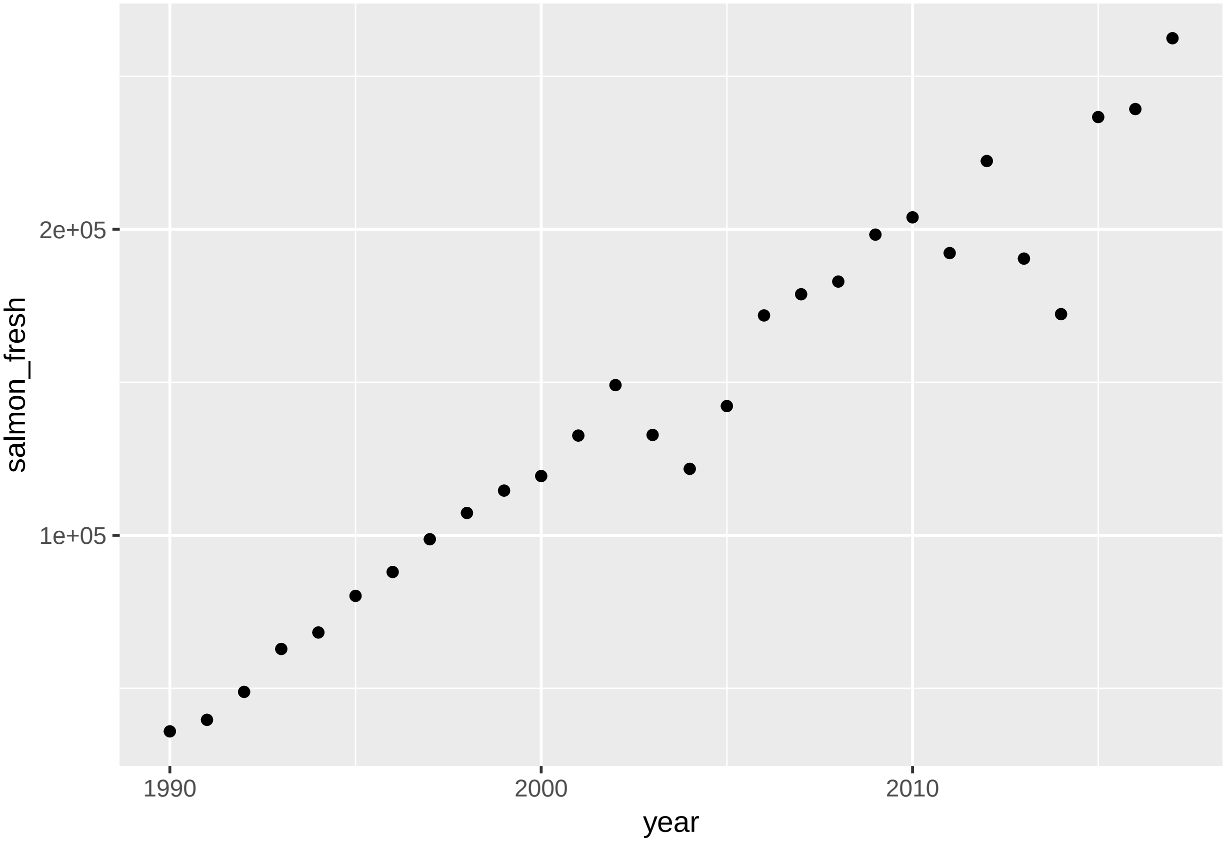
11/26/2018

Lab 8 - Linear Regression in R

```
library(tidyverse)
salmon <- read_csv("salmon_imports.csv")
```

```
salmon_scatter <- ggplot(salmon, aes(x = year, y = salmon_fresh)) +
  geom_point()

salmon_scatter
```



```
#Data look to be more or less linear.
```

```
#model_name <- lm(y_variable ~ x_variable, data = df_name)

salmon_lr <- lm(salmon_fresh ~ year, data = salmon)

salmon_lr
```

```
##
## Call:
## lm(formula = salmon_fresh ~ year, data = salmon)
```

```
##
## Coefficients:
## (Intercept)         year
##   -14982940         7550
```

Intercept is -14982940. Year is 7550.

Model equation: **Import(tons) = 7550*year - 14982940**

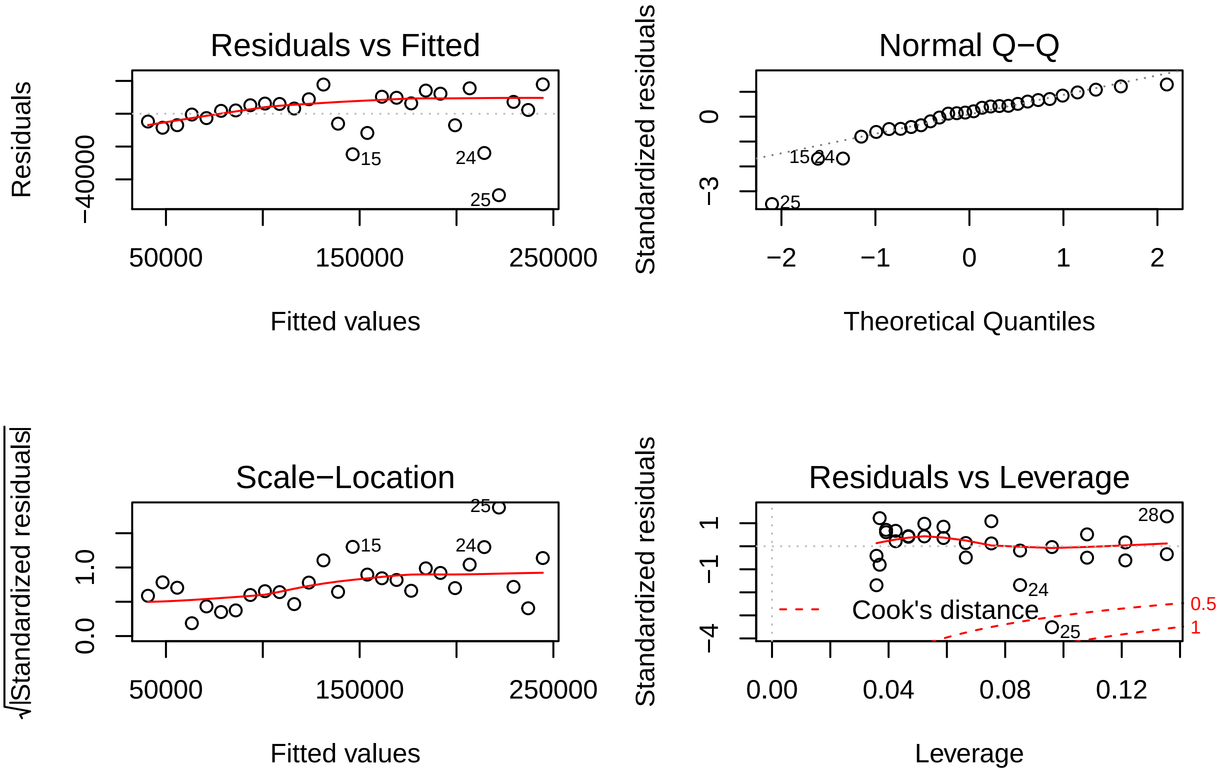
- What does the slope mean in the context of this model?

ANSWER: It meas that every year the amount of imported salmon increases by 7,550 tons.

- What does the y-intercept mean in the context of this model? Why is that concerning? What does this mean about extrapolating this model for past values?

ANSWER: It means that at year 0, we would be importing -14,982,940 tons of salmon. This is concerning because we can't be importing negative values (we're not considering that as exporting here). This means that we can't predict import values past a certain year in our dataset. Does this mean the model makes sense? Maybe yes, but just for a limited range or predictions going forward.

```r
par(mfrow = c(2,2))
plot(salmon_lr)
```



- Do residuals appear normally distributed?

ANSWER: They do not. The residuals appear to increase as the fitted values increase.

- Any concerns about heteroscedasticity or outliers?

ANSWER: Yes, it appears the data is heteroscedastic. We also have three outliers. Outlier 25 appears to be particularly far-out outlier.

```r
summary(salmon_lr)
```

```
##
## Call:
## lm(formula = salmon_fresh ~ year, data = salmon)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49619  -6284   2722   9063  17884
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.498e+07  6.963e+05  -21.52   <2e-16 ***
## year         7.550e+03  3.475e+02   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14860 on 26 degrees of freedom
## Multiple R-squared:  0.9478, Adjusted R-squared:  0.9458
## F-statistic: 471.9 on 1 and 26 DF,  p-value: < 2.2e-16
```

- Does year significantly predict salmon imports?

ANSWER: Yes.

- What does the R2 value actually mean in words?

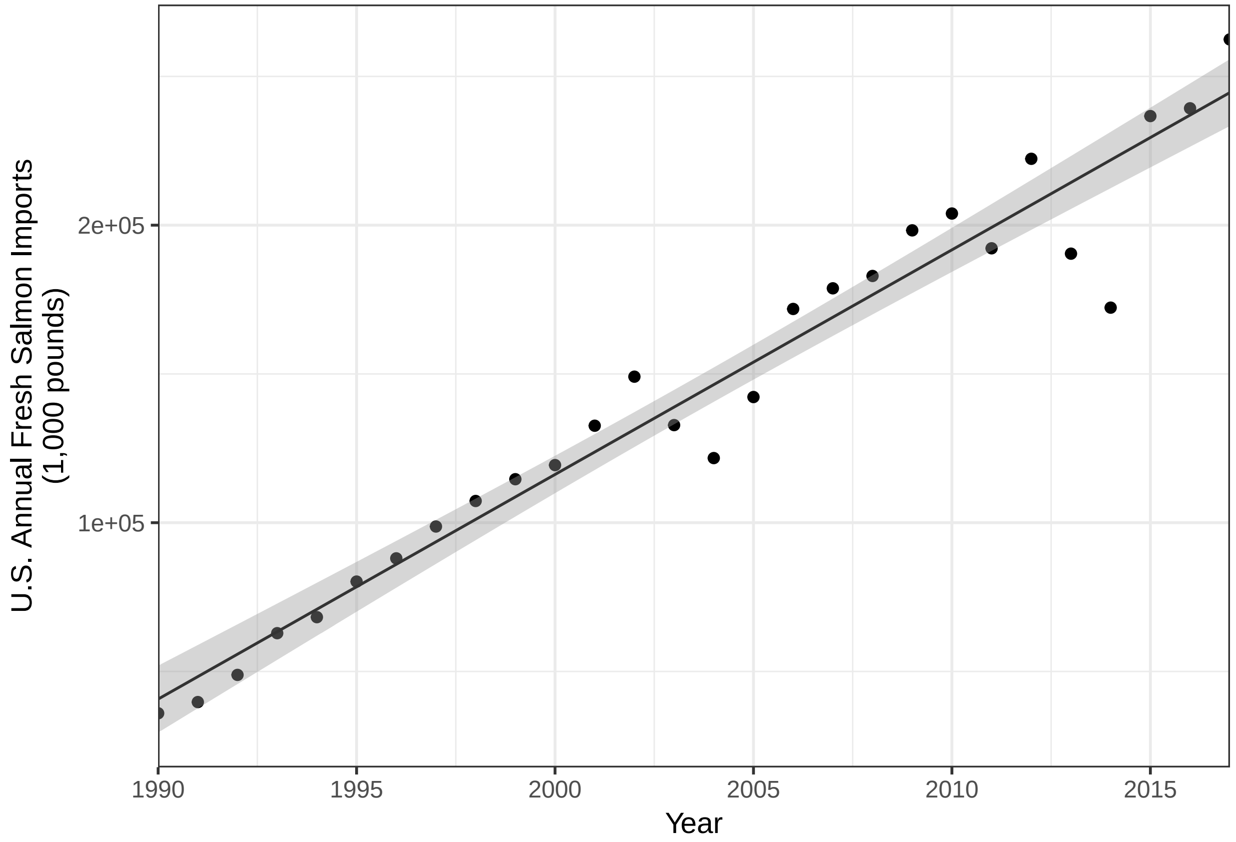
ANSWER: R2 descibes the proportion of variance in salmon imports (in tons) that is desribed by the year and the intercept (the model).

- What proportion of the variance in salmon imports is NOT explained by year?

ANSWER: 1-0.9458 = 0.0542 = ~5.4%.

```
salmon_final_graph <- ggplot(salmon, aes(x = year, y = salmon_fresh)) +
  geom_point() +
  geom_smooth(method = lm, se = TRUE, size = 0.5, color = "gray20") + #appeas as gray in output
  theme_bw() +
  scale_x_continuous(expand = c(0,0), limits = c(1990, 2017)) +
  labs(x = "Year", y = "U.S. Annual Fresh Salmon Imports\n(1,000 pounds)")

salmon_final_graph
```



```
new_years <- data.frame(year = c(2022, 2024, 2026, 2028))
new_years
```

```
##   year
## 1 2022
## 2 2024
## 3 2026
## 4 2028
```

```
future_predict <- predict(salmon_lr, newdata = new_years, interval = "confidence")
future_predict
```

```
##        fit      lwr      upr
## 1 282298.5 267877.4 296719.6
## 2 297397.6 281656.7 313138.5
## 3 312496.8 295418.5 329575.0
## 4 327595.9 309166.6 346025.2
```

```
predictions <- data.frame(new_years, future_predict)

predictions
```

```
##   year      fit      lwr      upr
## 1 2022 282298.5 267877.4 296719.6
## 2 2024 297397.6 281656.7 313138.5
## 3 2026 312496.8 295418.5 329575.0
## 4 2028 327595.9 309166.6 346025.2
```

```
salmon_cor <- cor(salmon$salmon_fresh, salmon$year)

salmon_cor
```

```
## [1] 0.9735387
```

```
#Value is 0.9735387
```

Pearson's r for the year vs. salmon imports linear trend:

ANSWER: 0.974

Would you describe this as a weak/strong negative/positive correlation?

ANSWER: Strong positive correlation.

Using the doc *Communicating Results of Basic Linear Regression* (posted on GauchoSpace) as a guide, write a final 1 - 2 sentence statement describing the results of your linear regression and Pearson's r findings.

ANSWER: