Multiple Linear Regression and Correlation

Lecture #7

Adi Sarid
Tel-Aviv University
updated: 2019-12-08

Reminder from previous lecture

We focused on simple linear regression.

- We saw how simple linear regression can be used to find the relationship between two variables
- For example, flight height and the number of bird strikes (or `log` or bird strikes)

We discussed the base assumptions in linear regression:

- Linearity $Y = \beta_0 + \beta_1 x + \epsilon$
- $\epsilon \sim \mathrm{N}(0, \sigma_\epsilon)$
- For hypothesis testing on β we also require homoscedastity

We discussed the objective function: the least squares L

We have shown how to find β_0 and β_1 from the partial derivative of $\partial L / \partial \beta_i$

Reminder from previous lecture (2)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Where:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$$

We discussed the importance of $SS_E = \sum_{i=1}^{n}(y_i - \hat{y})^2$, and talked about its role in estimating σ_ϵ:

$$\hat{\sigma}_\epsilon^2 = \frac{SS_E}{(n-2)}$$

The variance of the coefficients is given by:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{S_{xx}}$$

$$\text{Var}(\hat{\beta}_0) = \sigma_\epsilon^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

Reminder from previous lecture (3)

The variance of β_0, β_1 helped us devise a statistic and a hypothesis test for the parameters, i.e.:

$$T_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}_\epsilon^2 / S_{xx}}}$$

We've also seen the decomposition of the overall variance to the regression variance and error variance, i.e.:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2$$

$$SS_T = SS_R + SS_E$$

Where SS_R has 1 degree of freedome, SS_E has $n-2$ degrees of freedom, and SS_T has $n-1$ degrees of freedom.

Under a null hypothesis of $H_0 : \beta_1 = 0$, both SS_R, SS_E are χ^2 distributed, with $1, n-2$ degrees of freedom respectively.

This led us to an additional test, using analysis of variance.

Analysis of Variance for Regression Significance

Then, the following statistic would be F-distributed, under the null hypothesis:

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E}$$

The intuition behind the statistic is:

- As the mean square error MS_E decreases; and
- The variance explained by the regression model MS_R increases
- The model is a good fit to the data
- Hence, the null hypothesis of no model, i.e., $\beta_1 = 0$, is rejected

ANOVA (Analysis of Variance) Table

Source of Variation	**Sum of Squares**	**df**	**Mean Squares**	F_0
Regression	SS_R	1	MS_R	$\frac{MS_R}{MS_E}$
Error	SS_E	$n-2$	MS_E	
Total	SS_T	$n-1$		

Coefficient of determination R^2

We would like to measure the effect size of the regression. One possibility to measure the effect size is to use R^2:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Since $SS_T = SS_R + SS_E$, and all sizes are non negative:

$$0 \leq R^2 \leq 1$$

As the fit is better, R^2 increases.

Correlation

In probability, the correlation coefficient between two variables X and Y is defined as:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Where $\sigma_{XY}^2 = \text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_Y)] = E[XY] - E[X]E[Y]$

The correlation $\rho \in [-1, 1]$. When $\rho = 1 \text{ or } -1$, this means that the two variables have a linear relationship between them, and if it is 0 then the covariance is 0 and the two variables are independent.

We would like to see the relationship between ρ and R^2.

We can estimate ρ using:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Correlation (2)

We have defined $SS_T = \sum(y_i - \bar{y})^2 = S_{yy}$

Remember that:

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \hat{\rho}\sqrt{\frac{S_{yy}}{S_{xx}}}$, which demonstrates the relationship between $\hat{\rho}$ and β_1.

Also note that:

$$SS_E = \sum(y_i - \hat{y}_i)^2 = S_{yy} - \beta_1 S_{xy}$$

To see this use the above formulas for β_0 and β_1:

$$SS_E = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum\Big((y_i - \bar{y}) - \beta_1(x_i - \bar{x})\Big)^2 = S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx} = S_{yy} - \beta_1 S_{xy}$$

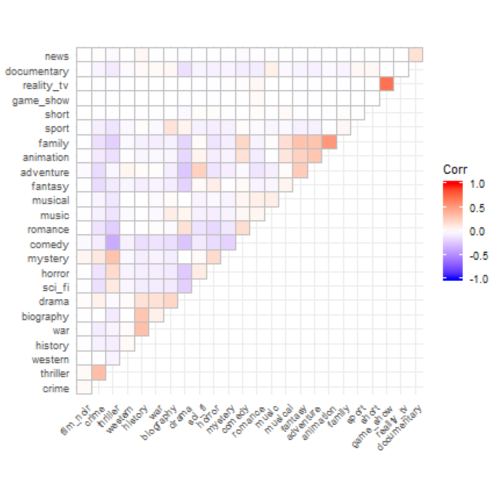
$$1 - R^2 = 1 - SS_R/SS_T = SS_E/S_{yy} = SS_E/SS_T = 1 - \beta_1 S_{xy}/S_{yy} = 1 - \frac{S_{xy}^2}{S_{xx}S_{yy}} = 1 - \hat{\rho}^2$$

Correlation (3)

Hence, we have proven that coefficient of determination R^2 is in fact the estimate for the square of the correlation coefficient $\hat{\rho}^2$ of X and Y.

In general, correlation is interesting because it can help us find simple association rules between variables. Let's see an example with movies genres. Adopted from a [scraped imdb](#) source.

```r
library(tidyverse)
# "https://raw.githubusercontent.com/sarid-ins/statistical_learning_course/master/datasets/scraped_imdb/movie
movies <- suppressWarnings(read_csv("data/movie_db_clean.csv", col_types = cols()) %>%
  janitor::clean_names() %>% select(adventure:western) %>%  mutate_all(~.*1))
genres_corr <- cor(movies)
#ggcorrplot::ggcorrplot(genres_corr, type = "upper", hc.order = T, hc.method = "complete")
```



Multiple Linear Regression - Background

So far we treated regression with only two variable (one dependent, `(Y)` and one independent `(X)`).

In most cases, we will have more than one independent variable, i.e., $X_1, \ldots, X_p$. Our model becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

This can also be extended to accomodate for more complex relationships such as:

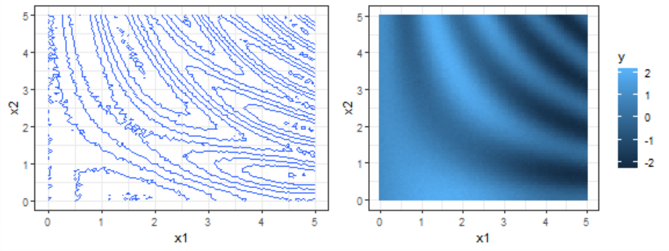
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon$$

or

$$Y = \sin(X_1) + \cos(X_1 X_2) + \epsilon$$

The principles and assumptions we discussed still hold, but some careful handling is needed, and this is what we will discuss today.

```r
ex1 <- crossing(x1 = seq(0, 5, 0.05),
         x2 = seq(0, 5, 0.05)) %>%
  mutate(y = sin(x1) + cos(x2*x1) + rnorm(NROW(x1), sd = 0.05))
p1 <- ggplot(ex1, aes(x = x1, y = x2, z = y)) +
  geom_contour() +
  theme_bw()
p2 <- ggplot(ex1, aes(x = x1, y = x2, fill = y)) +
  geom_tile() +
  theme_bw()
#cowplot::plot_grid(p1, p2)
```



Motivation

Let's analyze the example from last lecture (planes and birds), only this time, enrich the problem.

```
wildlife_medium <- read_csv("data/wildlife_impacts_medium.csv", col_types = cols())
lm(formula = log10(n) ~ ., data = wildlife_medium) %>%
  summary()
```

```
##
## Call:
## lm(formula = log10(n) ~ ., data = wildlife_medium)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77372 -0.39233 -0.02053  0.36176  2.02827
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.4689987  0.1360094  18.153  < 2e-16 ***
## skyOvercast    -0.3204574  0.0601218  -5.330 1.33e-07 ***
## skySome Cloud  -0.1103055  0.0516954  -2.134   0.0332 *
## rounded_height -0.0909040  0.0068071 -13.354  < 2e-16 ***
## rounded_speed   0.0002722  0.0004548   0.599   0.5496
## num_engs       -0.5461361  0.0431463 -12.658  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5961 on 698 degrees of freedom
## Multiple R-squared:  0.3501,    Adjusted R-squared:  0.3454
## F-statistic:  75.2 on 5 and 698 DF,  p-value: < 2.2e-16
```

Can you say what's the problem of examining this dataset, in this context? (think about possible biases)

Least Squares Estimation of the Parameters

We will use matrix notation. For each observation i we have:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

This can be represented as:

$$y = X\beta + \epsilon$$

Where:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We are looking for β which minimizes $L = \epsilon^t \epsilon = (y - X\beta)(y - X\beta)$

$$\frac{\partial L}{\partial \beta} = 0$$

Least Squares Estimation of the Parameters (2)

The resulting equations are given by:

$$X^t X \hat{\beta} = X^t y$$

In case that $X^t X$ is a non-singular matrix (i.e., invertible), the solution is unique and equals

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

Once $\hat{\beta}$ is found, we can use it to predict our values:

$$\hat{y} = X\hat{\beta}$$

We can also compute the residuals:

$$e = y - \hat{y}$$

Let $p = k + 1$ (the number of parameters including the constant β_0), then:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - p} = \frac{SS_E}{n - p}$$

Is an unbiased estimate of σ_ϵ^2

Hypothesis Tests

The vector $\hat{\beta}$ is an unbiased estimate:

$$E\hat{\beta} = E[(X^tX)^{-1}X^ty] = E[(X^tX)^{-1}X^t(X\beta + \epsilon)] = E[I\beta + (X^tX)^{-1}X^t\epsilon] = \beta$$

The β coefficients' variance is given by diagonal elements of $(X^tX)^{-1}$ times σ^2.

Now that we have found the expected value and the variance, we are ready for some hypothesis tests.

We are going to use the following set of hypothesis:

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$
- $H_1 : \exists i \text{ such that } \beta_i \neq 0$

Under the null hypothesis, SS_R/σ^2 is $\chi^2_{\mathrm{df}=k}$, and SS_E/σ^2 is $\chi^2_{\mathrm{df}=n-k-1}$.

Our statistic is:

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}$$

We reject H_0 if the computed value of the statistic $f_0 > f_{1-\alpha,k,n-k}$ (the right tail of F distribution)

Hypothesis Tests - ANOVA table

The process is summarized in an analysis of variance table, as follows:

Source of Variation	Sum of Squares	df	Mean Squares	F_0
Regression	SS_R	k	MS_R	$\frac{MS_R}{MS_E}$
Error	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Adjusted R^2

We discussed $R^2 = 1 - \frac{SS_E}{SS_T}$ however, as the number of parameters increases, the error always decreases, and the R^2 increases. This is prone to over-fitting (demonstration coming up).

To mitigate this phenomena we adjust the R^2.

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

Some Formulas for SS_E, SS_R, SS_T

Some formulas that we will use later on today. However, in practice the computer will calculate all the sizes we need.

$$SS_E = \sum (y_i - \hat{y}_i)^2 = e^t e$$

Substitute $e = y - \hat{y} = y - X\hat{\beta}$ we obtain:

$$SS_E = e^t e = (y - X\hat{\beta})^t (y - X\beta) = y^t y - 2\hat{\beta}^t X^t y + \hat{\beta}^t (X^t X)\hat{\beta} = y^t y - 2\hat{\beta}^t X^t y + \beta^t (X^t X)(X^t X)^{-1} X^t y = y^t y - \hat{\beta}^t X^t y$$

Since $SS_T = \sum (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2 / n = y^t y - (\sum y_i)^2 / n$ we get:

$$y^t y - (\sum y_i)^2 / n = SS_T = SS_R + SS_E = SS_R + y^t y - \hat{\beta}^t X^t y$$

Hence,

$$SS_R = \hat{\beta}^t X^t y - \left(\sum_{i=1}^{n} y_i \right)^2 / n$$

Remember these for later!

Overfitting a Regression Model: R^2 and adjusted R^2

Overfitting is a phenomena which occurrs when the number of parameters, p is very large compared to n

- In such a case a model is able to fit very well on the data
- On new observations, the model will suffer large errors

Various methods exist to handle and avoid this phenomena, such as train/test splits.

How would you devise an experiment to demonstrate overfitting?

See demonstration here.

03:00

Hypothesis Tests on Individual Coefficients

Sometimes, we are interested in the significance of a specific variable. I.e.,

- $H_0 : \beta_j = 0$
- $H_1 : \beta_j \neq 0$

First, remember that we noted that the j-th element on the diagonal of $\sigma^2(X^t X)^{-1}$ contains the variance of of the $\hat{\beta}_j$ (for an intuitive explanation see here).

Set $C = (X^t X)^{-1}$ then under the null hypothesis, we have the following student's-t statistic:

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}$$

The null hypothesis is rejected when $|t_0| > t_{1-\alpha/2, n-p}$

Hypothesis Testing - Subset of Coefficients

We now want to generalize the previous approach, which treated the null hypothsis of one variable into an arbitrary **subset** of variables.

Let $\vec{\beta}_1$ be a coefficient vector of $(r \times 1)$ and $\vec{\beta}_2$ a coefficient vector of $[(p-r) \times 1]$, i.e.:

$$\beta = \begin{bmatrix} \vec{\beta}_1 \\ \vec{\beta}_2 \end{bmatrix}$$

We want to test the hypothesis:

- $H_0 : \vec{\beta}_1 = \vec{0}$
- $H_1 : \vec{\beta}_1 \neq \vec{0}$

This yields two models:

The reduced null model: $y = X_2\vec{\beta}_2 + \epsilon \quad$ (the $\vec{\beta}_1$ variables are all 0).

The alternative full model $y = X\beta + \epsilon$

Hypothesis Testing - Subset of Coefficients (2)

Compute $SS_R(\cdot)$ for each of the models (the full model, the reduced model), also compute the contribution of the addition of $\vec{\beta}_1$:

$$SS_R(\beta) = \hat{\beta}^t X^t y, \quad \text{with } p = k + 1 \text{ degrees of freedom}$$

$$SS_R(\vec{\beta}_2) = \vec{\beta}_2^t X_2^t y, \quad \text{with } p - r \text{ degrees of freedom}$$

The regression sum of squares due to $\vec{\beta}_1$ given that $\vec{\beta}_2$ is already in the model, is:

$$SS_R(\vec{\beta}_1 \mid \vec{\beta}_2) = SS_R(\beta) - SS_R(\vec{\beta}_2), \quad \text{with } r \text{ degrees of freedom}$$

Our test statistic uses the partial F test (with $df = r, n - p$), i.e.,

$$F_0 = \frac{SS_R(\vec{\beta}_1 \mid \vec{\beta}_2)/r}{MS_E}$$

Reject the null hypothesis if $f_0 > f_{1-\alpha, r, n-p}$ (the right tail of the F-distribution).

Later we will discuss the stepwise algorithm, which aims to find good $\{\beta_i\}$ subsets, for improving the model

Hypothesis Testing - Subset of Coefficients - Example

Demonstration for a subset hypothesis test:

```
full_model <- lm(formula = log10(n) ~ ., data = wildlife_medium)
partial_model <- lm(formula = log10(n) ~ rounded_height + num_engs, data = wildlife_medium) # omitted sky, rounded_speed
SS_T <- sum((log10(wildlife_medium$n) - mean(log10(wildlife_medium$n)))^2)

aov_full <- anova(full_model)
aov_part <- anova(partial_model)

SS_R_full <- sum(aov_full$`Sum Sq`[1:4])
SS_E_full <- SS_T - SS_R_full
SS_R_part <- sum(aov_part$`Sum Sq`[1:2])
SS_R_add <- SS_R_full - SS_R_part

F_0 <- (SS_R_add/3)/(SS_E_full/(698))

df(F_0, df1 = 3, df2=698)
```

```
## [1] 4.509414e-06
```

```
qf(0.95, df1 = 3, df2 = 698)
```

```
## [1] 2.617663
```

We reject the null hypothesis with a p-value=4.509414×10^{-6}.

Confidence Intervals

For a coefficient's confidence interval, we can use the statistic:

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

Therefore, the two sided confidence interval is:

$$\hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

```
deg_f <- NROW(wildlife_medium) - 6
t_val <- qt(p = 0.975, df = deg_f)
full_model_sum <- summary(full_model)
ci <- tibble(coef = full_model$coefficients) %>%
   mutate(lb = coef - t_val*full_model_sum$coefficients[,2],
          ub = coef + t_val*full_model_sum$coefficients[,2] )
ci
```

```
## # A tibble: 6 x 3
##        coef        lb       ub
##       <dbl>     <dbl>    <dbl>
## 1  2.47      2.20      2.74
## 2 -0.320    -0.438    -0.202
## 3 -0.110    -0.212    -0.00881
## 4 -0.0909   -0.104    -0.0775
## 5  0.000272 -0.000621  0.00117
## 6 -0.546    -0.631    -0.461
```

Mean Response Confidence Interval

The point estimate for a new response at a point $x_0 = [1, x_{01}, \dots, x_{0k}]^t$ is:

$$\hat{\mu}_{Y|x_0} = x_0^t \hat{\beta}$$

The estimator is unbiased and its variance is:

$$\mathrm{Var}(\hat{\mu}_{Y|x_0}) = \sigma^2 x_0^t (X^t X)^{-1} x_0$$

Hence, we can use the following statistic for building our confidence interval:

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0}}$$

$$\hat{\mu}_{Y|x_0} + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0} \leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t_{1-\alpha/2, n-p} \sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0}$$

Prediction Intervals for New Observations

In case of a prediction interval for a new observation, the point estimate remains the same:

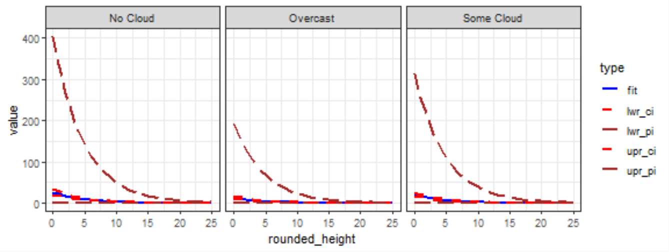
$\hat{y}_0 = x_0^t \hat{\beta}$

And the prediction interval is given by:

$$\hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + x_0^t(X^tX)^{-1}x_0)} \leq Y_0 \leq \hat{y}_0 + t_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + x_0^t(X^tX)^{-1}x_0)}$$

Example: prediction and mean response confidence intervals

```r
wildlife_new <- crossing(sky = c("No Cloud", "Some Cloud", "Overcast"),
                         rounded_height = 0:25,
                         num_engs = 2,
                         rounded_speed = 200)
new_responses_ci <- predict(full_model, newdata = wildlife_new, interval = "confidence") %>% as_tibble() %>%
   rename_at(vars(2:3), ~{paste0(., "_ci")})
new_responses_pi <- predict(full_model, newdata = wildlife_new, interval = "prediction") %>% as_tibble() %>%
   select(-fit) %>%
   rename_all(~{paste0(., "_pi")})
wildlife_tib <- wildlife_new %>%
   bind_cols(new_responses_ci,
             new_responses_pi) %>%
   mutate_at(vars(fit:upr_pi), ~10^(.)) %>%
   pivot_longer(cols = fit:upr_pi, names_to = "type", values_to = "value")
ggplot(wildlife_tib, aes(x = rounded_height, y = value, color = type, linetype = type)) +
   geom_line(size = 1) +
   facet_wrap(~sky) +
   scale_color_manual(values = c("blue", "red", "brown", "red", "brown")) +
   scale_linetype_manual(values = c(1, 2, 5, 2, 5)) +
   theme_bw()
```



Note About Extrapolation - Thought Experiment

What do you think is the problem with trying to provide an extrapolation (fit) and intervals (confidence for mean and prediction for a new observation) for the number of bird strikes with the following parameters:

- Flight height = 22 thousand feet
- Flight speed = 42 kts
- Sky = "No Cloud"
- Number of engines = 2

Can you think of a similar example but from a different domain?

03:00

Example - Outliers' Influence

Another "danger" in linear regression is what happens when the data contains outliers. Linear regression is very sensitive in this sense.

```
# wildlife_impacts <- readr::read_csv("https://raw.githubusercontent.com/rfordatascienc
# write_csv(wildlife_impacts %>% count(height), "lectures/data/wildlife_impacts_small.c
wildlife_small <- read_csv("data/wildlife_impacts_small.csv", col_types = cols()) %>%
  mutate(rounded_height = round(height/1000)) %>%
    group_by(rounded_height) %>%
    summarize(n = sum(n)) %>%
    filter(!is.na(rounded_height))

wildlife_err <- wildlife_small
wildlife_err[19, 2] <- 600000 # instead of 6 we multiplied this observation by 1000

p1 <- ggplot(wildlife_small, aes(x = rounded_height, y = log10(n))) +
    geom_point() +
    stat_smooth(method = "lm") + coord_cartesian(ylim = c(-1, 5)) + theme_bw()
```

How are Types of Variables Used in Regression?

As you have probably noticed, in the bird-planes example, we used a `sky` variable which has three values (factor). The regression model is linear, if so, how are factor variables treated?

- Factors are turned into dummy variables (0/1).
    - How many dummies are needed for a 3-level factor? why?
- Characters are treated the same
- Ordinals - depending on definition, might be entered as polynomials, factors, or continuous
- Logicals - as a 0/1 variable

Questions:

- What is the meaning of the coefficient β of a logical variable?
- What is the meaning of the coefficient β of a factor?
- How would you consider a date type variable?

03:00