# Simple Linear Regression and Correlation

## Lecture #6

Adi Sarid
Tel-Aviv University
updated: 2019-11-30

# Reminder from previous lecture

We talked about two sample statistics

- Hypothesis testing of means, e.g.: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$
- Confidence intervals, e.g.: $\mu_1 - \mu_2 \in \bar{x}_1 - \bar{x}_2 + [z_{\alpha/2}, z_{1-\alpha/2}]\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
- Power computations desired sample size with two sample statistics
- Paired versus unpaired tests
- Comparing variance using the $F$ test
- Goodness of fit for examining variable independence, i.e., using $\chi^2$ test
- We saw some examples (men vs. women, mobile phone and health)

# Inference on two population proportions

We consider the case of two binomial parameters $p_1, p_2$. Let $X_1, X_2$ represent the number of successes in each sample. $\hat{P}_i = X_i/n_i$, have approximately normal distributions.

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Is distributed approximately as $Z \sim \mathrm{N}(0, 1)$.

Under the null hypothesis $H_0 : p_1 = p_2 = p$ we have:

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}}$$

Where an estimator to $p$ is given by:

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2}$$

# The test procedure for comparing two population proportions

Null hypothesis: $H_0 : p_1 = p_2$

Test statistic: $Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\,(1/n_1 + 1/n_2)}}$

Alternative hypothesis (rejection criteria):

- $H_1 : p_1 \neq p_2 \quad (z_0 > z_{1-\alpha/2}$ or $z_0 < z_{\alpha/2})$
- $H_1 : p_1 > p_2 \quad (z_0 > z_{1-\alpha})$
- $H_1 : p_1 < p_2 \quad (z_0 < z_{\alpha})$

# Setting the sample sizes when comparing two population proportions

Very similar to what we've shown in the last lecture for one sample, but with a slightly different computation for the standard deviation under $H_1$. For example, in the two sided case we have:

$$\beta = \Phi\left[\frac{z_{1-\alpha/2}\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}}\right] - \Phi\left[\frac{z_{\alpha/2}\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}}\right]$$

With $\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$, $\bar{q} = \frac{n_1(1-p_1) + n_2(1-p_2)}{n_1 + n_2}$ and

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

We can obtain the suggested sample (or power) using `pwr::pwr.2p.test` or `pwr::pwr2p2n.test`.

```
pwr::pwr.2p2n.test(h = pwr::ES.h(p1 = 0.2, p2 = 0.3),
                   n1 = 150, n2 = NULL,
                   sig.level = 0.05,
                   power = 0.8,
                   alternative = "less")
```

# Effect Size

We discussed p-value as the extent to which a statistical finding is significant. However, it is not the sole measure for the strength of a statistical finding.

In this context, see the ASA statement on $p$-Values here

**Effect size** measures the magnitude of a phenomena. Effect size is a generic name for various measures such as:

- $R^2$ in linear regression
- $\rho$ Pearson correlation coefficient between two variables
- Cohen's $d$ which relates to the difference between means (which we will now discuss)
- Many more

# Effect Size - Cohen's $d$

The difference between two means divided by standard deviation, i.e.:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S_p}$$

Where $S_p$ is the pooled standard deviation:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

```
set.seed(0)
ipf <- read_csv("data/ipf_lifts.csv", col_types = cols()) %>%
   filter(best3squat_kg > 0) %>%
   sample_n(1000)

effsize::cohen.d(formula = best3squat_kg ~ sex, data = ipf)
```

```
##
## Cohen's d
##
## d estimate: -1.863577 (large)
## 95 percent confidence interval:
##     lower     upper
```

# Simple Linear Regression: background

Linear regression is an important modeling technique in statistics. It was developed in the 1800s and is still very common today.

- In essense, it allows us to model the relationship between two or more variables, utilizing some basic assumptions such as linearity, and normality.
- These days, it is less common as a predictive modeling approach, because for predictions we have much better models.
- It is still heavily used in research (e.g., academia) to describe and indicate statistically significant relationships.
- Linear regression is very appealing as one of the "first models to try out" because it is very simple to understand, has a low computational price, it is easy to interpret, and yet very flexible.

# Simple Linear Regression: example - bird strikes (1/3)

A very troubling problem for aviation is bird strikes

- From a monetary perspective - causing damages to planes
- From a safety perspective - endangers the passanges and crew
- (Obviously it's not that fun to the birds either)

What is the relationship between flight height and the number of bird strike events?

The data we will be exploring is adopted from tidytuesday (2019-07-23), here.

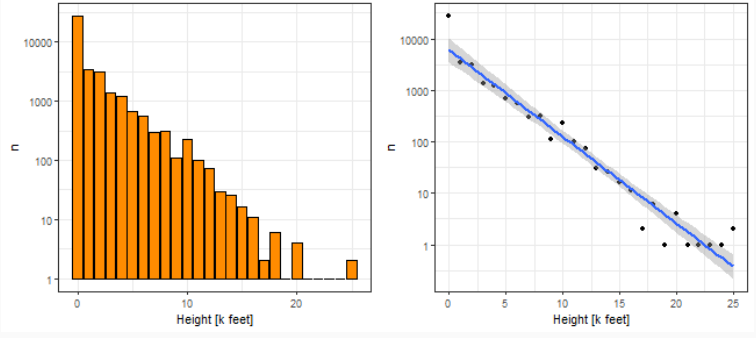
```
# wildlife_impacts <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2019/2019-07-23/wildlife_impacts.csv")
# write_csv(wildlife_impacts %>% count(height), "lectures/data/wildlife_impacts_small.csv")
wildlife_small <- read_csv("data/wildlife_impacts_small.csv", col_types = cols()) %>%
  mutate(rounded_height = round(height/1000)) %>%
   group_by(rounded_height) %>%
   summarize(n = sum(n)) %>%
   filter(!is.na(rounded_height))
```

# Bird strike events example (2/3)

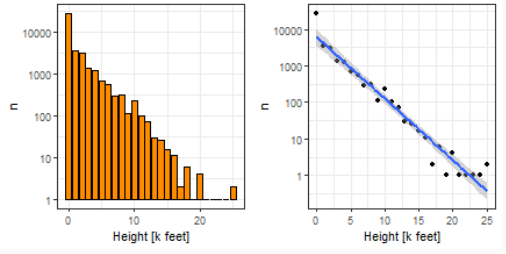
The was categorized to intervals of 1000 feet, i.e., $0 - 999, 1000 - 1999, \ldots, 25000$.

Note that the data y-axis appears in **log-scale**.

```
wildlife_hist <- ggplot(wildlife_small, aes(x = rounded_height, y = n)) +
   geom_col(fill = "darkorange", color = "black") + scale_y_log10() + theme_bw() + xlab("Height [k feet]")
wildlife_points <- ggplot(wildlife_small, aes(x = rounded_height, y = n)) +
   geom_point() + scale_y_log10() + theme_bw() +
   stat_smooth(method = "lm") + xlab("Height [k feet]")
cowplot::plot_grid(wildlife_hist, wildlife_points)
```



# Bird strike events example (3/3)




It would seem as though each additional $5k$ feet decrease the number of bird strikes by a ratio of $10$, or in other words:

$$\log(\text{Bird strikes}) \approx 3784 - 0.168 \times h$$

Equivalently, we can also write:

$$\text{Bird strikes} \approx 10^{3784 - 0.168 \times h}$$

Even thought this is not exactly a linear equation, it was obtained using linear regression, and we will see later on how we reached this formula.

# The Basic Regression Model: description and assumptions

At the base of linear regression, we describe the relationship between an independent variable $Y$ and a dependent variable $X$ as:

$$E(Y|X = x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

The regression coefficient $\beta_0$ is called the **intercept**, and $\beta_1$ is called the **slope** (why?).

This relationship is generalized as:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

Where $\epsilon$ is assumed to be distributed as $\mathrm{N}(0, \sigma_\epsilon)$.

This model is called a **simple linear regression model**

- Only one independent variable (only a single $x$, aka regressor)

Note that:

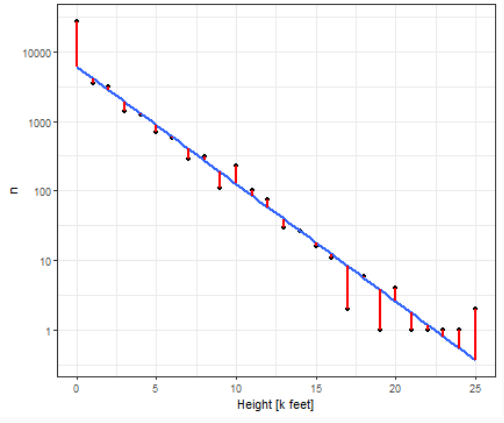
$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E\epsilon = \beta_0 + \beta_1 x$$

$$\mathrm{Var}(Y|x) = \mathrm{Var}(\beta_0 + \beta_1 x) + \mathrm{Var}(\epsilon) = \sigma_\epsilon^2$$

# Properties of the least squares estimators

The most common method to find the linear relationship is called the least squares estimate. I.e., we are looking for the line which brings to minimum the suqared errors. I.e.:

- The $\min \sum_i (\hat{y}_i - y_i)^2$ of the red lines in :



# Finding the coefficients using the least squares method

For each observation $i$, we have: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, \ldots, n$

The sum of squares is given by

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

To find the estimators for $\beta_0, \beta_1$ require:

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0$$

# Finding the coefficients using the least squares method (2)

Simplifying the equations we obtain:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

and

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

The solution to these equations is given by:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - \frac{\sum_i y_i \sum_i x_i}{n}}{\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}}$$

The fitted line is then $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

# Finding the coefficients using the least squares method (3)

Set

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$S_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

Then:

$$\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

# The residuals

Each observation satisfies $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, 2, \ldots, n$. We call $e_i$ the $i$th residual.

We define $SS_E$, the error sum of squares, as:

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

To estimate $\sigma_\epsilon^2$ we can use $SS_E$:

$$E(SS_E) = (n-2)\sigma_\epsilon^2$$

An unbiased estimator for $\sigma_\epsilon^2$ is therefore:

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

# Linear Regression - Demonstration via R (1)

We'll now demonstrate how to run linear regression in R, and then continue the discussion about linear regression.

```
mtcars_lm <- lm(formula = mpg ~ disp, data = mtcars)
summary(mtcars_lm)
```

```
##
## Call:
## lm(formula = mpg ~ disp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8922 -2.2022 -0.9631  1.6272  7.2305
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.599855   1.229720  24.070  < 2e-16 ***
## disp        -0.041215   0.004712  -8.747 9.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.251 on 30 degrees of freedom
## Multiple R-squared:  0.7183,    Adjusted R-squared:  0.709
## F-statistic: 76.51 on 1 and 30 DF,  p-value: 9.38e-10
```

# Linear Regression - Demonstration via R (2)

In this example we also use a transformation within the formula of `lm`, i.e. `log(n)`

```
wildlife_lm <- lm(formula = log(n) ~ rounded_height, data = wildlife_small)
summary(wildlife_lm)
```

```
##
## Call:
## lm(formula = log(n) ~ rounded_height, data = wildlife_small)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4258 -0.2640 -0.0495  0.2009  1.6772
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.71284    0.25826   33.74   <2e-16 ***
## rounded_height -0.38788    0.01772  -21.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6775 on 24 degrees of freedom
## Multiple R-squared:  0.9523,    Adjusted R-squared:  0.9503
## F-statistic: 479.3 on 1 and 24 DF,  p-value: < 2.2e-16
```

# Properties of the Least Squares Estimators

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are a result of a computation based on our sample, i.e., $\bar{y}, \bar{x}, S_{xx}, S_{xy}$.

Specifically, they depend on the observed $y$'s and hence they are random variables themselves. Both coefficients are unbiased, i.e., $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$.

The variance of the coefficients is given by:

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

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

The covariance of $\beta_0$ and $\beta_1$ is given by:

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = E(\hat{\beta}_0 - E\hat{\beta}_0)(\hat{\beta}_1 - E\hat{\beta}_1) = -\sigma_\epsilon^2 \frac{\bar{x}}{S_{xx}}$$

# Hypothesis tests in simple linear regression

Now that we found a relationship $y = \hat{\beta}_0 + \hat{\beta}_1 x$, we want to figure out, is this relationship "real"?

In other words, we would like to test the hypothesis:

- $H_0 : \beta_0 = 0$
- $H_1 : \beta_0 \neq 0$

And the hypothesis:

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

If both are rejected (especially $\beta_1$), we can say that the relationship we found is statistically significant.

To develop the statistical test for the coefficients we will need the following assumptions:

- The errors are normally distributed, i.e. $\epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon)$, and
- The errors are homoscedastic, i.e., no matter the $x_i$, the error distribution is the same

# Hypothesis tests in simple linear regression (2)

Recall that

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - \frac{\sum_i y_i \sum_i x_i}{n}}{\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}}$$

Since $Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma_\epsilon^2)$, this means that $\hat{\beta}_1$ is also normally distributed (as a linear combination of normal variables).

$$\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma_\epsilon^2 / S_{xx})$$

The following statistic is distributed student's t, with $n - 2$ degrees of freedom

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

That is, reject $H_0$ if:

$$t_0 > t_{1-\alpha/2, n-2} \quad \text{or} \quad t_0 < t_{\alpha/2, n-2}$$

# Hypothesis tests in simple linear regression (3)

A similar test can be used for $\beta_0$: $T_0 = \frac{\hat{\beta}_0 - 0}{\sqrt{\hat{\sigma}_\epsilon^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}}$

Now, we can have another look on the linear regression output.

```
summary(wildlife_lm)
```

```
##
## Call:
## lm(formula = log(n) ~ rounded_height, data = wildlife_small)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4258 -0.2640 -0.0495  0.2009  1.6772
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.71284    0.25826   33.74   <2e-16 ***
## rounded_height -0.38788    0.01772  -21.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6775 on 24 degrees of freedom
## Multiple R-squared:  0.9523,    Adjusted R-squared:  0.9503
## F-statistic: 479.3 on 1 and 24 DF,  p-value: < 2.2e-16
```

# Analysis of Variance for Regression Significance

So far we treated the coefficients individually, however, we want a different hypothesis test which will examine the regression as a whole.

The variance of the $y_i$ observations can be broken in the following manner:

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

The variability of the data (the LHS) is broken down to the variability explained using the regression line, and the remaining variability (the errors $e_i^2$).

This is sometimes noted as

$$SS_T = SS_R + SS_E$$

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.

The expectancy of each element is $E[SS_E/(n-2)] = \sigma_\epsilon^2$, $E[SS_R] = \sigma_\epsilon^2 + \beta_1^2 S_{xx}$.

Under a null hypothesis of $H_0 : \beta_1 = 0$ we obtain that both sum of squares are $\chi^2$ distributed.

# Analysis of Variance for Regression Significance (2)

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** | $\boldsymbol{F_0}$ |
|---|---|---|---|---|
| Regression | $SS_R$ | $1$ | $MS_R$ | $\frac{MS_R}{MS_E}$ |
| Error | $SS_E$ | $n-2$ | $MS_E$ | |
| Total | $SS_T$ | $n-1$ | | |

# Confidence intervals

As with any parameter, we can compute confidence intervals for $\beta_0$ and $\beta_1$:

$$\beta_1 \in \hat{\beta}_1 \pm t_{\alpha/2,n-1}\sqrt{\frac{\sigma_\epsilon^2}{S_{xx}}}$$

$$\beta_0 \in \hat{\beta}_0 \pm t_{\alpha/2,n-2}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

# Prediction intervals

We would like to use the linear regression model for prediction of new values.

Given $x_0$, our prediction for $Y_0$ is

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

This is a point estimate, however, we are also interested in a **prediction interval**, i.e., where would a new observation lie in probability of 95%.

Observe that the error of the new observation is: $e_p = Y_0 - \hat{Y}_0$.

The mean of the error is 0 and the variance of the error is:

$$\mathrm{Var}(e_p) = \mathrm{Var}(Y_0 - \hat{Y}_0) = \sigma_\epsilon^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

Hence:

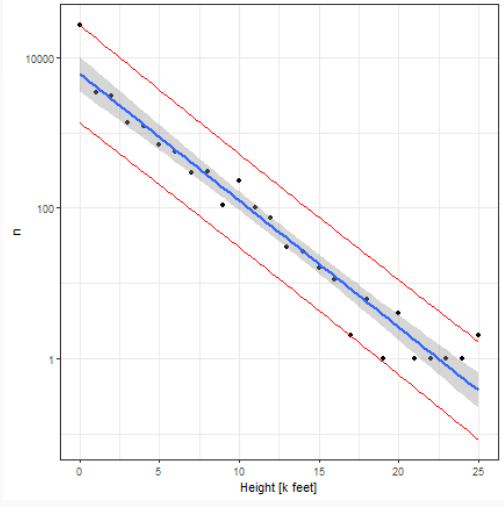
$$\sqrt{\sigma^2 \left[ \quad 1 \quad (x_0 - \bar{x})^2 \right]}$$

# Prediction intervals - illustration

The grey band shows confidence intervals around $Y_0 | X = x$ and the outer lines show the prediction intervals

```
wildlife_predicted <- predict(wildlife_lm,
                              newdata =
                                 tibble(rounded_height = seq(0, 25, by = 0.25)),
                              interval = "prediction") %>%
    data.frame() %>%
    mutate(rounded_height = seq(0, 25, by = 0.25))
```

```
ggplot(wildlife_small, aes(x = rounded_height, y = n)) +
   geom_point() + theme_bw() +
   xlab("Height [k feet]") +
   geom_line(data = wildlife_predicted,
             inherit.aes = T, aes(x = rounded_height, y = exp(lwr))) +
   geom_line(data = wildlife_predicted,
             inherit.aes = T, aes(x = rounded_height, y = exp(upr))) +
   scale_y_log10() +
   stat_smooth(method = "lm")
```

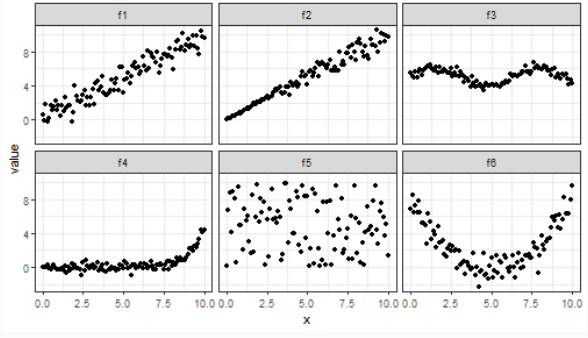


# Would linear regression fit here?

We mentioned these assumptions to linear regression:

- Linear relationship
- Normal error $\mathrm{N}(0, \sigma_\epsilon^2)$
- Homoscedastity

Which of the following violates these assumptions? Can you think of a transformation that would fix the problem?




05:00

# 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.

Next week we will also dive a bit deeper into transformations and how to utilize them, and talk about correlation between variables.

We will talk about multiple linear regression, and discuss some cavets of $R^2$, of overfitting, and how to overcome them.